Evaluation of Web Search Engines

by

Yali Wang
B.Eng, Tianjin University, 1998

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of

MASTER OF APPLIED SCIENCE

in the Department of Electrical and Computer Engineering

Supervisory Committee

Evaluation of Web Search Engines

by

Yali Wang
B.Eng, Tianjin University, 1998

Supervisory Committee

Dr. Kin F. Li, (Department of Electrical and Computer Engineering)
Supervisor

Dr. Xiaodai Dong, (Department of Electrical and Computer Engineering)
Departmental Member

Dr. Jens Weber, (Department of Computer Science)
Outside Member

Dr. Kui Wu, (Department of Computer Science)
External Member

Abstract

Supervisory Committee

Dr. Kin F. Li, (Department of Electrical and Computer Engineering)
Supervisor

Dr. Xiaodai Dong, (Department of Electrical and Computer Engineering)
Departmental Member

Dr. Jens Weber, (Department of Computer Science)
Outside Member

Dr. Kui Wu, (Department of Computer Science)
External Member

Using the proper search engine is crucial for efficient and effective web search. The objective of this thesis is to develop methodologies to evaluate search engines in a systematic and reliable manner. A new model for evaluation and comparison of search engines is proposed. This hierarchical model classifies the most common features found in search engines and search results into groups and subgroups. To illustrate the usefulness of the proposed model, several Chinese search engines are evaluated and compared as a case study. It is also very important to evaluate the performance of a search engine over time. Three performance measurement metrics are formulated for this purpose. Performance results for English and Chinese search engines are represented by histograms for visual inspection. The histograms are classified into groups to facilitate the interpretation of the performance metrics and examination of the associated behaviours of the search engines. An automated classification method is developed that eliminates the subjectivity and ambiguity found in visual classification of the histograms.

Table of Contents

List of Tables

List of Figures

Acknowledgments

I would like to express my heartiest appreciation to my supervisor Dr. Kin Li for his continuous guidance and encouragement shown throughout this research work and the process of writing this thesis. I cannot thank you enough for your patience and forbearance with me.

I wish to express my gratitude to Z. Alam for his help and support during my research. Without his help, I could not have finished this thesis. I also wish to thank my aunty Yunxia and my cousin Lily for their continuous help and support. Thanks to all my friends, especially Lynn, Xiao, Wei, and my Chachi, who made me feel that Victoria is my second home.

Finally I want to thank my parents for their continuous care and support.

Chapter 1

Introduction

Since its inception, the Internet has rapidly expanded in size and complexity. The World Wide Web (hereafter refers to as the web in this work) now consists of billions of web pages in many different languages. Web search engines have become indispensable tools for the users. A search engine has to search through a massive amount of information and to provide users the most relevant results in a reasonable short time. The challenge in the design of novel web search engines has generated a great deal of interest in both industry and academia.

1.1 Internet Search Engines

A search engine is the software used to retrieve information from a database or from the Internet. Among the search engines available today, two basic designs are employed: directory based and query based. Yahoo is one of the best known directory-based search engines and it also has the ability to perform query-based search. Other commonly used search engines, such as Google, AltaVista, and HotBot, have adopted the query based approach.

Directory-based search engines feature a hierarchically organized subject tree, and often the tree is formulated by humans. Users can either traverse the subject index via links, or search against the index using a simple keyword query. Entries in the index are created manually by human reviewers. This is unlike in systems such as Google where internal indices are created automatically and searched via queries. There are two ways to locate pages to index. One is based on submissions by individuals and corporations to the search engine provider, and the other is using an automated system that searches the web and returns appropriate documents.

Query-based search engines have become more popular. In this research, the focus is on query-based search engines. A typical query-based search engine consists of spiders or crawlers, web collection, indexer, index, and query analyzer as shown in

Figure 1.1 [12]. In order to provide efficient search performance, a search engine generally includes features to enhance its search capability, such as Boolean operators, search fields, search modifiers, etc.

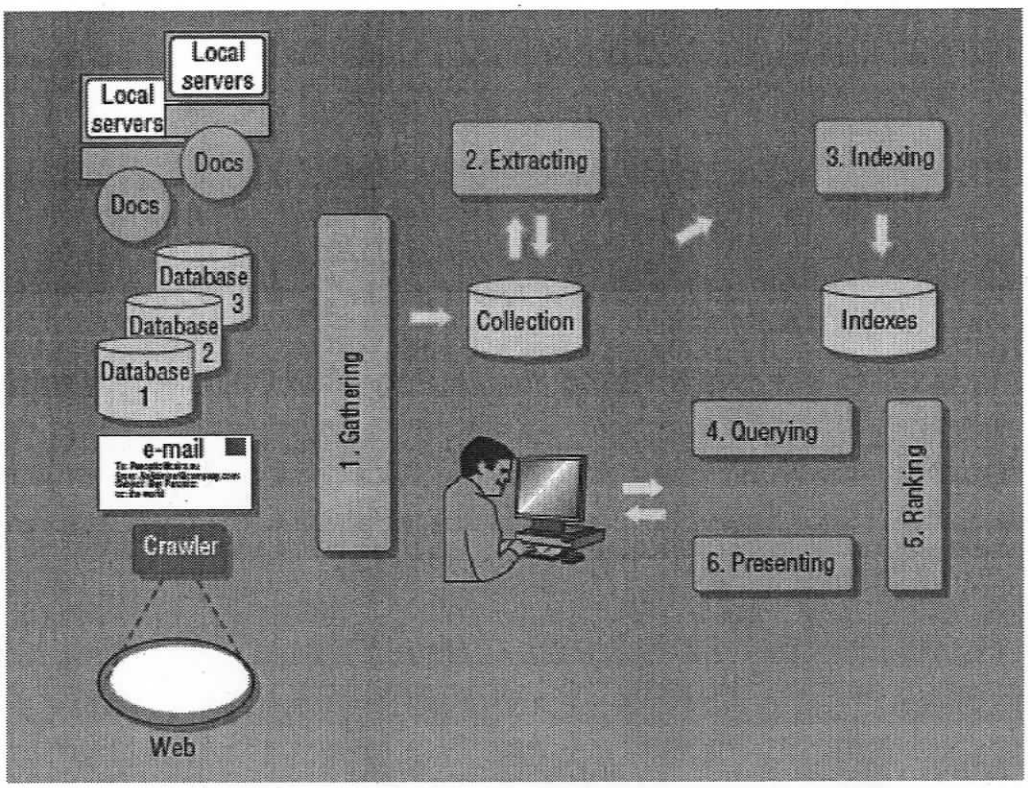


Figure 1.1 *General structure of a query-based search engine*

The spiders collect information from the web and save it in the web collection. An indexer is used to index the information tracked by the spiders. The indices are saved in an index database. When a user wants to search information using a search engine, he/she would input keywords in the user interface. The query analyzer then goes to the index database to find the most relevant results that match the keywords, and presents them to the user.

1.2 Motivation in the Evaluation of Search Engines

A survey shows that around 85% of the Internet users employ search engines to find information on the web [34]. Finding the right information in a short time with the least

effort is very important for both corporate and individual users. An efficient search engine could mean a great saving of money, time and effort. Evaluating the search engines and finding the appropriate ones to use is therefore of great importance.

When searching the web, most people rely on a few search engines only. For searching in English, Google, Yahoo and MSN are the most widely used search engines. The performance of a search engine depends on many issues, such as coverage of the databases, index strategies, query features, ranking algorithm, etc. Currently, Google dominates the search engine world; however there is no clear indication or proof that Goggle is superior to other search engines in every way. Internet users are a very diverse group with very different preferences and priorities. For some users, finding up-to-date information in an easy manner may be most important. For others, finding in-depth analyses may be more important than finding recent information. It is therefore very difficult to select one single search engine that all users would find satisfactory.

Developing models and methods to evaluate search engines in a reliable and consistent manner can assist users to choose the most suitable search engine serving their purposes. This could result in great savings of money and resources. The evaluation can also provide insight into the characteristics of the search results and therefore is useful to the search engine providers to develop and improve their engines.

1.3 Objectives and Contributions

The objective of this research work is to develop methodologies to evaluate search engines in an easy and reliable manner. An evaluation model for search engines is proposed. In this model, the most common features found in search engines are classified into groups and subgroups. These feature parameters are given scores according to the specified criteria, thus enabling the evaluation of a search engine. To illustrate the usefulness of the proposed model, an evaluation of several Chinese search engines is used as a case study.

It is important to evaluate the performance of a search engine over time. Three performance measurement metrics are formulated and the results are represented by histograms for visual examination. Data from both English and Chinese search engines are collected over several time periods. The patterns of the histograms generated from the

collected data show some similar properties. The histograms are classified into groups for easy interpretation of the performance metrics and the associated behaviour of the search engines. In addition, an automated classification method is developed that eliminates the subjectivity and ambiguity found in visual classification by humans.

1.4 Outline of This Thesis

The rest of this thesis is organized as follows.

Chapter 2 presents background information on search engine evaluation and existing work in the literature. Factors that influence a search engine's performance are also addressed.

Chapter 3 introduces a model for search engine evaluation. Characteristics of search engines and properties of search results are classified into groups and subgroups for further in-depth evaluation and comparison.

Chapter 4 presents a case study of the proposed model: the evaluation of several major Chinese search engines. The characteristics of Chinese search engines are also discussed.

Chapter 5 investigates the variation of search results from different search engines over time. Three performance measurement metrics are formulated. Data collected for English and Chinese search engines are presented using histograms. Classification of the resulting histograms is discussed. A deterministic classification scheme is introduced.

Chapter 6 summarizes the contributions of this thesis and suggests some possible future work.

Chapter 2

An Introduction to Search Engine Evaluation

With the explosive growth of web, the importance of and the interest in search engine evaluation have greatly increased in recent years. This chapter aims at providing the relevant background information necessary in order to follow the work presented in later chapters. The importance of search engine evaluation is discussed and the evaluation criteria are introduced. Relevant works performed by other groups are reviewed.

2.1 The Importance of Search Engine Evaluation

With the rapid growth of the web, users demand to find the right information from the web efficiently and effectively, with minimum effort. This phenomenon has inspired the development of search engine technology. There are numerous search engines available satisfying both general-purpose and specialized requirements [34] [50]. Therefore, selecting the most appropriate search engine for their particular search can result in saving a great deal of time and effort for the users. In addition, research on search engine evaluation can assist the search engine providers to gain insight to their products and improve the quality of service. Many search engines claim to use novel information retrieval (IR) techniques and their evaluation is also of interest to the IR research community. Search engine evaluation is therefore becoming a very important field of research.

2.2 Search Engine Evaluation Criteria

For any evaluation method, the natural first step is to choose the proper evaluation criteria. Since search engine evaluation is a rather new area of research, there are still debates on the proper evaluation criteria to be used.

The web is in fact a huge database and a search engine is a special type of IR system. Therefore, it is reasonable to use the same evaluation criteria to compare search

engines as used in IR systems [18] [19] [48]. Lancaster and Fayen [17] listed six criteria for the evaluation of an IR system: coverage, recall, precision, response time, user effort and form of output. Additional evaluation criteria were introduced by others. Cooper et al. proposed the use of expected search length [58], novelty ratio and relative recall [31]. Some researchers combined precision and recall in a single measure [56].

Search engines, however, differ significantly from that of other IR systems in terms control. The web has no rules and no central administrator. This can be seen as one of the great advantages of the web. On the other hand, this has made information retrieval from the web a formidable task. Traditional IR systems are highly controlled, centralized and operated in a relatively stable environment. The web on the other hand is uncontrolled, distributed and highly dynamic. Chakrabarti et al. [47] rightly commented that "the Web has developed into a global mess of previously unimagined proportions". The changes on the web are continuous and the search engines are not always aware of them. The search engine has to search billions of web pages and gather the information from this chaotic environment. The relevant number of pages can be enormous but the user usually spends very little time to go through the retrieved pages. Silverstein et al. found that about 85 percent of the users look at the results on the first page only [9]. These challenges have made the working principles of search engines and their evaluation criteria significantly different from those of the traditional IR systems.

There is no unanimous agreement on the evaluation criteria for search engines. Various research groups have used different criteria for ranking search engines. Some of the most widely used criteria are described here.

Precision

Precision or precision ratio P of a search engine is defined as [37]

$$P = \frac{D_r}{D_t}$$

where D_r is the number of relevant documents retrieved and D_t is the total number of retrieved documents. Relevance or precision is somewhat subjective. Two searchers can look for the same information on the same topic. The third-ranked document may appear relevant and useful to searcher A but not useful at all to searcher B. In spite of

subjectivity and inconsistency, precision is still one of the most important criteria for search engine evaluation.

Average precision [37] is a common measure used to assess the retrieval performance. Let r_i denote the number of relevant documents up to and including position i in the returned list of documents. The recall at the i^{th} document in the list (assuming the first document is most relevant) seen thus far is $R_i = \frac{r_i}{r_n}$, where r_n *is* the total number of relevant documents in the collection. The precision at the i^{th} document, P_i is defined as the proportion of documents up to and including position i that are relevant to the given query. The pseudo-precision at a recall level x is then defined as

$$\tilde{P}(x) = \max(P_i) \tag{2.1}$$

where $x \le \frac{r_i}{r_n}$ and $i=1, 2, \ldots, n$

Using Eq. (2.1), the n point interpolated average precision for a query is given by

$$P_{av} = \frac{1}{n}\sum_{i=0}^{n-1} \tilde{P}\left(\frac{i}{n-1}\right) \tag{2.2}$$

As it is common to observe retrieval at recall levels k/10, for k=0, 1, …, 10, an n=11 point average precision (P_{av}) is typically used to measure the performance of information retrieval for each query. If a single measure is desired for multiple queries, the mean or median P_{av} across all queries can be used.

Recall

Recall or recall ratio R of a search engine is defined as [37]

$$R = \frac{D_r}{N_r} \tag{2.3}$$

where D_r is the number of retrieved relevant documents and N_r is the total number of relevant documents in the collection. Recall ratio is difficult to obtain since the total number of relevant documents is very often unknown. Nevertheless, recall remains a very useful concept.

There is always a tradeoff between precision and recall [40]. If a searcher wants only the precise documents that fit his/her exact needs, the query will require very

specific terms. However, in that case, there is a risk of missing relevant documents. Therefore, the search must be broadened to include more relevant documents in the search result. On the other hand, the precision of the broadened search will drop, and the user has to wade through many irrelevant documents. Figure 2.1 illustrates this point more clearly.

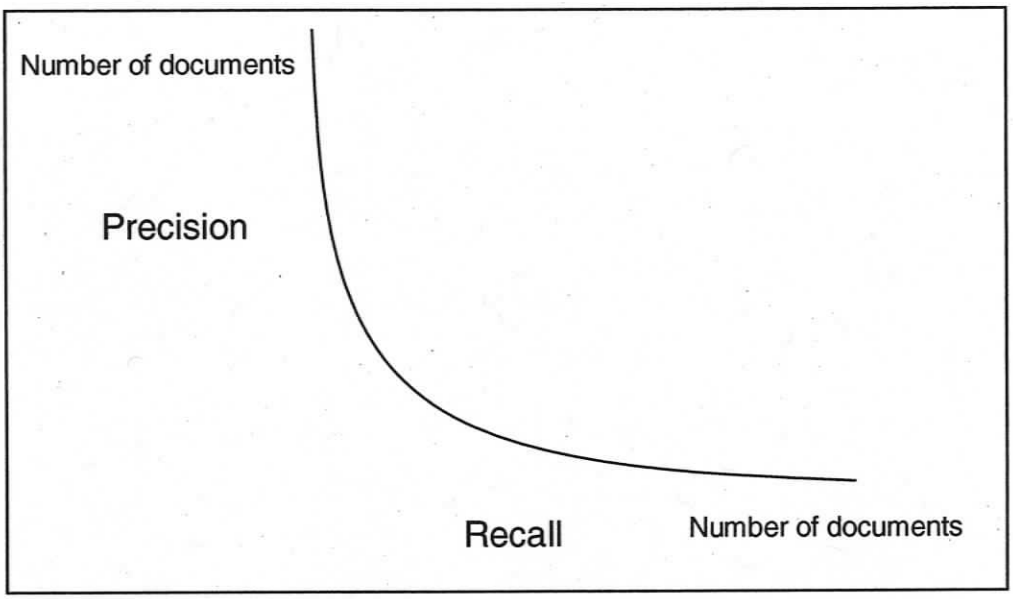


Figure 2.1 *Number of document distributions for a given query showing the inverse relationship between Precision and Recall*

From the figure, it can be observed that the larger the number of documents is retrieved (Recall), the fewer the number of those documents is relevant (Precision).

Coverage

The web is already huge in size and is growing at an enormous pace everyday. Estimates from multiple research groups in April 2005 put the indexed web at 11.5 billion pages with other estimates citing an additional 500-plus billion non-indexed and invisible web pages yet to be indexed [14]. Estimates in April 2005 from Google.com, Yahoo.com, Cyberatlas and MIT[14] had 45 billion web pages being publicly available and another 5 billion pages being available within private intranet sites.

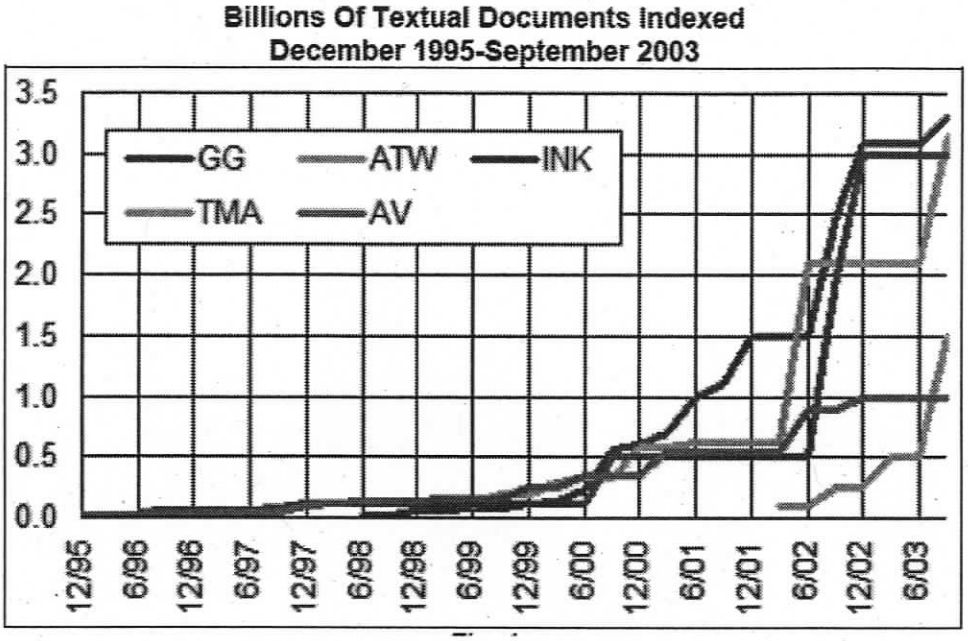


Figure 2.2 *Billions of textual documents indexed by different search engines*

Key : GG = Google ATW = AllTheWeb INK = Inktomi (now Yahoo!) TMA = Teoma (not Ask Jeeves) AV = Alta Vista (now Yahoo!) Source: Search Engine Watch, January 28, 2005.

The number of web pages indexed by search engines is increasing rapidly. Figure 2.2 [14] shows the increasing rate in the number of web pages indexed by some of the major search engines. There is a considerable amount of the web that is not indexed or covered by any search engine. Gulli [3] estimated the visible web (uniform resource locators or URLs the search engines can reach) to be more than 11.5 billion pages. Table 2.1 [3] presents the web coverage by some of the top search engines in 2005.

Search Engine	Self-Reported Size (Billions)	Estimated Size (Billions)	Coverage of Indexed Web (%)	Coverage of Total Web (%)
Google	8.1	8.0	76.2	69.6
Yahoo	4.2 (est.)	6.6	69.3	57.4
Ask	2.5	5.3	57.6	46.1
MSN (beta)	5.0	5.1	61.9	44.3
Indexed Web	N/A	9.4	N/A	N/A
Total Web	N/A	11.5	N/A	N/A
Note: "Indexed Web" refers to the part of the web considered to have been indexed by search engines.				

Table 2.1 *Web coverage of some major search engines*

Relative coverage is more commonly used for search engine comparison. When comparing a number of search engines, relative coverage is defined as the total number of relevant URLs returned by a search engine divided by the total number of URLs returned by all search engines [22]. This measure has two major limitations. First, even if all the databases of all search engines are merged, it does not cover the entire web. According to a study in 2000, 8% of the web is composed of disconnected components, which are very difficult to reach by the crawlers [1]. Second, several URLs can be textually different but point (or redirect) to the same physical location. As a result, the number of relevant URLs returned by a search engine may be an overestimate. Nevertheless, relative coverage is still widely used to compare the coverage of different search engines.

User Effort

In contrast to traditional IR searchers, the majority of the web users are laypersons who are more sensitive to the time and effort spent on finding information [4] [9]. Korfhage [45] pointed out that users usually do not tolerate more than three or four attempts of feeding back information to the system. The ability to optimize search order thus becomes an even more salient dimension of search engine performance. The notion of Expected Search Length (ESL), first proposed by Cooper (1968) some 30 years ago, seems to be ideal to test how well a search engine is able to deliver the most relevant documents at the top of the retrieved sets [53], [29], [8], [33]. According to Cooper, the primary function of a retrieval system is to save users as much labour as possible in the search for relevant documents by perusing and discarding irrelevant ones.

Form of Output

The interface of a search engine is a critical issue from the user's perspective. Since there are so many search engines available, the attractiveness of each search engine is expressed, to its users, to a great extent in its documentation and interface. Users may not feel interested to use a search engine unless they are comfortable with its interface, and are able to read and comprehend its documentation when necessary. This evaluation component should be examined from two perspectives. One is the number of output options a web search engine offers, whereas the other deals with the actual content of the

output. Sometimes, one search engine may appear quite impressive in one aspect, but it has weakness in other evaluation facets.

2.3 Review of Search Engine Evaluation

In literature, there are two approaches for search engine evaluation: testimonials and shootouts [32] [11] [23] [2] [60]. Testimonials are casual studies and state the general impression obtained after a few queries. Shootouts are rigorous studies and follow the information retrieval measures in the evaluation process.

The work done by Gordon and Pathak is considered as one of the earliest and most extensive work based on testimonial method [32]. Thirty-three volunteers participated in the experiment. The top 20 results generated by the search engines were then returned to the volunteers for relevance assessment. They found that the search effectiveness was low, there were significant differences between the engines, and the rankings of the engines were to some extent dependent on the strictness of the relevance criterion. They mentioned the following 7 desirable features of a web search evaluation process [32]:

1. Searches should be motivated by genuine user's need.
2. If a search intermediary is employed, the primary searcher's information need should be as fully captured as possible and transmitted in full to the intermediary.
3. A large number of search topics must be used.
4. Most major search engines should be included.
5. The most effective combination of specific features of each search engine should be exploited, i.e., the queries submitted to the engines need not be the same.
6. Relevance judgment must be made by the individual who needs the information.
7. Experiments should be well designed and conducted.

There are some disagreements about the importance of some of these features. Hawking et al. [11] pointed out that features 1 and 6 would limit the scope of the evaluation process. As for feature 5, they argued that public search engines are designed

to produce a list of results with a set of query words in the search box provided. Therefore, it makes more sense to compare the quality of search engine results by identical input queries.

Almost all evaluation studies involve human interaction and judgement, which is very time consuming and costly. More importantly, subjectivity makes a study less reliable. So it is highly desirable to automate the evaluation process. Chowdhury et al. [2] proposed one of the earliest automated methods. It is based on searching items in the Open Directory Project (OPT), and comparing their relative ranks in the search engines' returned lists. They compared five well known search engines (Lycos, Netscape, Fast, Google, HotBot) and concluded that the performance of the engines are statistically almost equivalent. Liu et al. [60] proposed click-through data analysis. They used the click-through data collected for a set of queries. The mean reciprocal rank is used as the evaluation criterion. Here reciprocal rank is the reciprocal of the correct answer's ranking in a search engine's result list. They compared five popular search engines (Baidu, Google, Yisou, Sina, Zhongsou) using this method. The performance of Baidu was the best and Google was also very satisfactory.

The majority of early works on search engine evaluation found in public literature do not stress the fact that search engine performance may change over time. The web is highly dynamic. New web pages and URLs are added continuously. Many web pages are removed resulting in dead links. A good search engine should be capable of reflecting these changes quickly in their search results. Bar-Ilan [23] has studied these features extensively and has proposed several criteria for search engine evaluation:

- Timeliness / freshness – A search engine should show less number of broken links and it should list even the most recently created web pages.
- Stability over time – the number of results in each query should be stable over time.

In most search engine evaluation schemes, data are usually collected over a period of time. A quick way to analyze the variation of the results over time is to represent the data in histograms or other kinds of graphs. However, analysis of data by visual inspection is challenging. Bar Ilan et al. [25] have proposed a method to handle this problem in an automated manner. They used a number of statistical measures to quantify the variation

of search results including Spearman's foot rule and Fagin's measure. They observed significant difference between the results obtained from different engines.

The performance of a search engine depends crucially on the ability of the user to provide the appropriate query. A search engine may be very suitable for an expert user to find some particular information but it may not be the best option for a novice. This point has been taken into account by some researchers recently [49]. They provided an economic model for comparing search services based on users' requirements. This model can help users to select the search service that minimizes cost, maximizes benefits and reduces uncertainty. The search services providers can also use the model to enhance their services and products.

2.4 Conclusion

In the past few years, search engine has become an indispensable tool for Internet users. Studies on search engine evaluation have become very useful. Still in its early stage of development, more research effort is needed to investigate search engine evaluation criteria as well as fully automated and reliable search engine evaluation techniques.

Chapter 3

A Search Engine Evaluation Model

People can find information very quickly using any one of the existing search engines by simply entering the desired keywords. The usefulness of the returned hits is open to question, and is left to be judged by the user. The 'goodness' of the results depends on the choice of the search words as well as the effectiveness of the search engine. There are many search engines available these days, though only a few dominate. Each search engine has its own characteristics and effectiveness depending on the user's keywords and search criteria. It is not easy for a user to choose the most appropriate search engine for his/her particular use. In chapter 2, a review on search engine evaluation is given. The most common criteria for evaluating search engines, such as precision, recall, user effort, and coverage, are introduced. Existing works in the literature are also described. Most search engine evaluation works do not include many criteria into their evaluation model. After analyses of existing evaluation models and search engine characteristics, seventy evaluation criteria, including some new ones, are categorized into two groups, features and performance. Detailed descriptions of these two criteria groups and our evaluation model are given in this chapter.

3.1 Overview of Our Evaluation Model

Most existing search engine comparisons either simply review a few factors, or focus on certain aspects of the search results, or rate several aspects of the site subjectively. In order to perform a thorough evaluation of the search engines and make more meaningful comparisons, we should explore and review as many different factors as possible. Seventy parameters are identified and used in our evaluation model including commonly used web metrics as presented in Section 3.3.

The seventy parameters are classified into two major groups based on their functionality. The "Performance Group" includes features and capabilities that enhance

the usability of the engine. The "Features group" includes various metrics for evaluating search results.

Elements in each group can be further subdivided into subgroups and sub-subgroups, resulting in a hierarchy of evaluation parameters. This provides additional flexibility to the users. Because of the clear hierarchical structure it is easy for users to focus on a specific group of evaluation parameters of interest. Users can also compare several search engines based on the selected features. This is accomplished by the use of a weighted sum of the parameters as presented in Section 3.2

To evaluate precision and recall of an IR system, the traditional way is to use public benchmarks such as TREC (Text Retrieval Conference) [15] and NTCIR (NII-NACSIS Test Collection for IR Systems) [39] for the experiments. Unfortunately, evaluating the efficiency and effectiveness of web search engines creates many unique challenges that make a TREC-style evaluation problematic. First, the web can be viewed as an unbounded database and therefore makes certain evaluation metrics meaningless. Second, the web is too large to have relevance judgement done by humans. Third, the web is 'live' and is changing continuously, and therefore requires special performance metrics to make the evaluation meaningful. In this research, we use the concept of a common list as a baseline for relevance evaluation. The common list is described in details in Section 3.4.2.

3.2 Weighed Parameters and Summary Score

A hierarchical structure allows us to rate the search engines at various abstraction levels of details. In general, the mathematical model or score of a collection of parameters can be expressed as

$$Score = \sum_{i=1}^{X} w_i P_i \tag{3.1}$$

where w_i is the weight assigned to the i^{th} parameter of a group with X parameters. Considering the two major groups of evaluation parameters, the total score of a search engine is represented as

$$Score = w_{feature} P_{feature} + w_{performance} P_{performance} \tag{3.2}$$

A very important part of developing the evaluation model is the assignment of weights to different parameters. The sign for the weight is rather objective. If the presence of a feature makes the search engine more useful, the sign is positive, otherwise it is negative. A negative weight can be used to indicate the undesirable impact of a parameter on the total score. For example, the higher the number of dead links in the returned results, the worse the search engine is compared to others. Therefore the parameter indicating the number of dead links should have a negative weight. On the contrary, the magnitude of a weight is subjective. If one feels that performance is more important than the features of a search engine, the weights assigned may be 0.6 and 0.4, respectively.

The scores for features and performance, $P_{feature}$ and $P_{performance}$, in turn, are derived from subsequent scoring equations at lower levels of the evaluation hierarchy. The value of a parameter is either a 0 or 1 to indicate whether a feature or capability exist in that search engine. A range between 0 and 1 is assigned to parameters that have various degrees of quality. The sum of the weights assigned to all parameters within a group must be equal to 1. This ensures the consistency of weight distribution among different groups.

This flexibility of tailoring the scoring system to individual needs, by changing the weights of the parameters or deleting an unwanted feature or adding a desirable feature, makes the proposed evaluation model very attractive to search engine users, subscribers, and providers. As pointed out in a workshop position paper [21], specific web search engines are effective only for some types of queries in certain contexts. Using our model and making the proper adjustment as described, a user will be able to find the particular search engine that suits his/her needs.

3.3 Feature Parameters

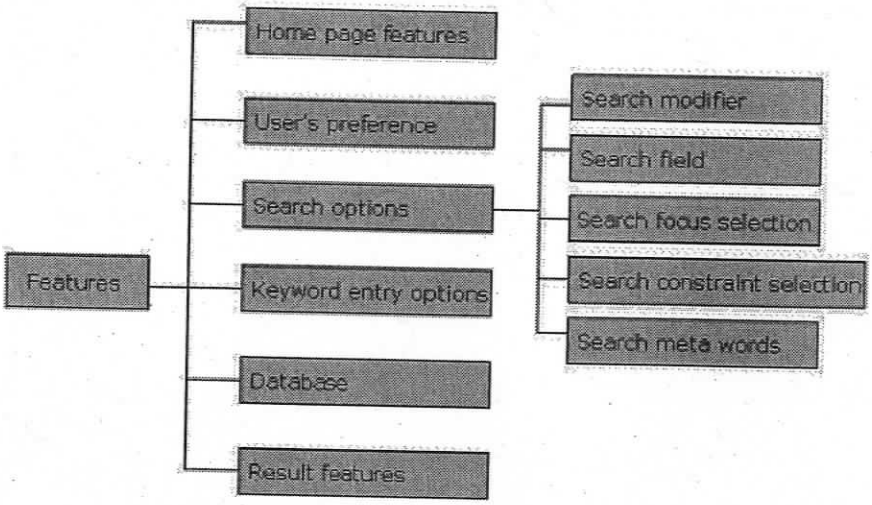


Figure 3.2 *Feature parameters considered*

We further classified feature parameters into six major categories as shown in Figure 3.1. This hierarchical structure pools collections of related parameters into subgroups and sub-subgroups:

1. Home Page Features: This category indicates how user friendly the home page is regarding various help and user selection menus. This feature includes a subjective user evaluation, the availability and visibility of help links, result language selection, topic directory selection, and advanced search selection.
2. User Preferences: This category includes a choice of the home page language, the availability of safe search filtering, the control of the number of results per page, the choice of displaying the results in a new window, intelligent input correction, search default setting, search options within the result page, and news search.
3. Search Options: This category is further divided into subgroups of search modifier (at least one, case sensitive, etc.), search field (title, url, links, etc.), search focus selection (web site, web page, directory, etc.), search constraint selection (language, file format, publication date, etc.), and search meta words for focused search (specified sites only, similar pages, geographic regions, etc.).

4. Keyword Entry Options: This category considers the capability of the search engine in stop word interpretation, case sensitivity, exact phrase specification, wildcard allowance, search by pronunciation, and Boolean operators.
5. Database: This category indicates the number of groupings arranged in directories and the total number of indexed pages of the corresponding search engine.
6. Result Features: This category reviews display features such as, whether there is indication for the total number of hits, the number of pages, and search time; the capability to search within results; whether the results are ordered and numbered; whether different file formats are allowed in the returned items; whether a pay listing is allowed (a negative weight); web page snap; further search for related pages and the presence of hits' date, size and summarization.

The following tables show the detailed information for each feature group including the description of each parameter, the possible value of the parameter, whether the evaluation is subjective or not (if users evaluate according to their own judgement, the evaluation is subjective; it is objective if the evaluation is based on facts.) , and the sign of the assigned weight.

Description	Parameter's value	Subjective	Sign of Weight
User preferences	0 to 1		
Homepage interface language selection	yes=1, no=0	No	"+"
Safe search filtering	yes=1, no=0	No	"+"
Number of results (number of results per page)	yes=1, no=0	No	"+"
Results window (in a new window)	yes=1, no=0	No	"+"
Intelligent input correction	yes=1, no=0	No	"+"
Set the search default homepage	yes=1, no=0	No	"+"
Search option on the result page	yes=1, no=0	No	"+"
News display capability on search result page	yes=1, no=0	No	"+"

Table 3.1 *Feature parameters: user preferences*

Description	Parameter's value	Subjective	Sign of Weight
Database			
Directory: normalized number of categories	0 to 1	No	"+"
Database: normalized total number of pages	0 to 1	No	"+"

Table 3.2 *Feature parameters: database*

Description	Parameter's value	Subjective	Sign of Weight
Homepage features	0 to 1		
User's evaluation = average (User1+User2+User3)	0 to1	**Yes**	"+"
Help link	yes=1, no=0	No	"+"
Result language selection	yes=1, no=0	No	"+"
Directory search selection	yes=1, no=0	No	"+"
Advanced search selection	yes=1, no=0	No	"+"

Table 3.3 *Feature parameters: homepage features*

Description	Parameter's value	Subjective	Sign of Weight
Results feature	0 to 1		
Total hits indication	yes=1, no=0	No	"+"
Number of pages indication	yes=1, no=0	No	"+"
Show search time	yes=1, no=0	No	"+"
Search within results	yes=1, no=0	No	"+"
Results are ordered	yes=1, no=0	No	"+"
Results are numbered	yes=1, no=0	No	"+"
Results in various file formats	yes=1, no=0	No	"+"
Paid listing/result	yes=1, no=0	No	"-"
Web page snap	yes=1, no=0	No	"+"
Related pages	yes=1, no=0	No	"+"
Check pages in the same hit's website	yes=1, no=0	No	"+"
Date of the hit	yes=1, no=0	No	"+"
Size of the hit	yes=1, no=0	No	"+"
Comments of the hits	yes=1, no=0	No	"+"

Table 3.4 *Feature parameters: results feature*

Description	Parameter's value	Subjective	Sign of Weight
Search options			
Search modifier	0 to 1		
Include all of the following keywords	yes=1, no=0	No	"+"
Include the following exact phrase	yes=1, no=0	No	"+"
Exclude	yes=1, no=0	No	"+"
At least one of	yes=1, no=0	No	"+"
Case sensitivity specification	yes=1, no=0	No	"+"
Others	yes=1, no=0	No	"+"
Search field (return results where the terms occur)	0 to 1		
Anywhere in the page	yes=1, no=0	No	"+"
In the title of the page	yes=1, no=0	No	"+"
In the text of the page	yes=1, no=0	No	"+"
In URL of the page	yes=1, no=0	No	"+"
In the links of the page	yes=1, no=0	No	"+"
Others	yes=1, no=0	No	"+"
Search focus selection			
Web site	yes=1, no=0	No	"+"
Web page	yes=1, no=0	No	"+"
Directory	yes=1, no=0	No	"+"
Others (e.g., mp3, news, images)	yes=1, no=0	No	"+"
Search constraint selection	0 to 1		
Result language selection	yes=1, no=0	No	"+"
Result file format	yes=1, no=0	No	"+"
Time limiting capability	yes=1, no=0	No	"+"
Search meta words	0 to 1		
Site: return results from the specified URL or domain	yes=1, no=0	No	"+"
Similar: return pages that are similar to the hit	yes=1, no=0	No	"+"
Links: find pages that link to the hit	yes=1, no=0	No	"+"
Geographic region: search within the selected ranges	yes=1, no=0	No	"+"
Others (e.g., inurl, intitle)	yes=1, no=0	No	"+"

Table 3.5 *Feature parameters: search options*

Description			Parameter's value	Subjective	Sign of Weight
Keyword entry options			0 to 1		
	Stop word		yes=1, no=0	No	"+"
	Case sensitivity		yes=1, no=0	No	"-"
	Exact phase (use quotation mark)		yes=1, no=0	No	"+"
	Asterisk wildcard (e.g. "*", "?")		yes=1, no=0	No	"+"
	Search by pronunciation		yes=1, no=0	No	"+"
	Boolean operators				
		AND/and/"+"/" "	yes=1, no=0	No	"+"
		OR/or/"|"/"/"	yes=1, no=0	No	"+"
		NOT/not/"-"	yes=1, no=0	No	"+"
	Others		yes=1, no=0	No	"+"

Table 3.6 *Feature parameters: keyword entry options*

3.4 Performance Parameters

3.4.1 Performance Groups

There are three major groups of performance metrics as shown in Figure 3.2: the response time, the total number of hits as indicated on the result page, and the quality of results.

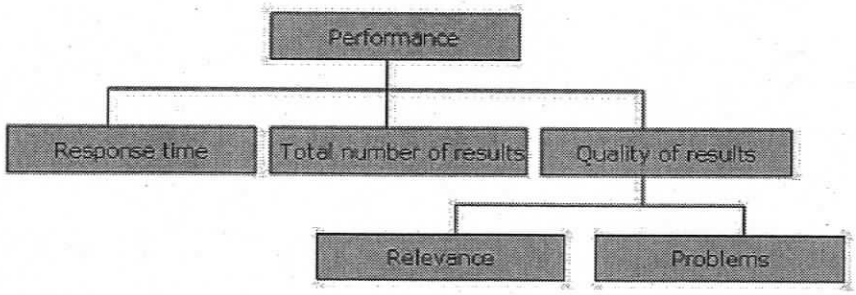


Figure 3.3 *Performance parameters*

The Quality of Results group consists of subgroups Problems and Relevance. The Problems subgroup indicates the severity and frequency of problems encountered when a

user tries to access the search engine or the returned hits. This includes the number of times that the search site is down during trial and experimentation, the number of broken links such as host not found and connection time out, and duplicates. All these parameters within the Problems group carry negative weights.

In order to obtain the Relevance score, one can solicit the assistance of humans to examine the relevance of the returned items with respect to the keywords. The scores are averaged as an attempt to eliminate any inherent potential bias and subjectivity in human interpretation. On the other hand, one can eliminate the subjectivity of humans by using a common list approach as described in Section 3.4.2. In such case, the Relevance subgroup includes the parameters precision @ N and recall @ M.

The following table shows the detailed information for each performance parameter including its description, its possible values, whether it is subjective or not, and the sign of its weight.

Description	Parameter's value	Subjective	Sign of Weight
Quality of the results	0 -- 1		
Relevance	0 -- 1	Yes	"+"
Precision @ N	yes=1, no=0	No	"+"
Recall @ M	yes=1, no=0	No	"+"
Problems	yes=1, no=0	No	"-"
Response time	Normalized number	No	"+"
Number of hits	Normalized number	No	"+"

Table 3.7 *Performance parameter: description, value, subjectivity and sign of the weight*

3.4.2 Evaluating Relevance Using a Common List

In order to eliminate subjectivity and the labour intensive process of using humans to provide the relevance score, an algorithmic approach is desirable. Comparing individual search engine's results against a common baseline is a reasonable means to evaluate relevance. The critical issue is how to generate such a common list. Since all search engines return what they regard as high ranking items in an ordered list, therefore it makes sense to combine the individual lists into a single list. This single list can be considered as the most accurate list as it consolidates the expertise of all the search

engines. Matching items to this common list gives a sense of quality as well as quantity in terms of relevance.

Traditionally, there are three metrics commonly used to measure the relevance of the matched results in information retrieval of a bounded database:

- Precision: the proportion of the returned items that are deemed relevant.
- Precision @ N: precision evaluated from the top N highest ranked items.
- Recall: the proportion of relevant items returned.

In the context of web search, it is impractical to examine all the returned items of a search engine, which is in the order of thousands and even hundreds of thousands. Also, it is impossible to measure the total of number of relevant items exists on the web, not to mention that this number is changing constantly. Therefore the first and the third metrics listed above are impossible to determine. We need to modify the definition of the above metrics to fit our current context with some assumptions.

First, an item in a search engine's list is deemed relevant if it also appears in the common list. This makes sense as it can be assumed that all the search engines are experts and the consensual common list derived collectively has a high probability of holding the truely high-ranked items. Second, extending this argument, we can treat the number of items in the common list as the total relevant items, since we are interested only in highly relevant items in most cases. Third, most search engine users are interested only in the top ten or so items as appear on the first result page, and therefore considering only the top ten ranked items is sufficient for our purpose.

For the Relevance factor, the following two revised metrics are used:

- *Precision @ N* = the number of items that also appear in the common list over *N*.
- *Recall @ M* = the number of items that also appear in the common list over the number of items *M* in the common list.

In many other relevance evaluations, precision @ N is taken as the ratio of relevant items as interpreted by humans over that of the top N items, while recall is not measured at all [18]. Our revised precision @ N concept eliminates human effort and subjectivity by matching items to a common list. These revised metrics actually measure relevance in terms of both quantity and quality. Two commonly employed metrics, coverage and

overlap, are not used in our study because these measures have already been indirectly incorporated into the common list.

3.5 More on Quality Issues

There are many different types of problems encountered when accessing a web page, for examples, dead link, page not found, site busy, etc. All these problems are related to how frequent a search engine re-visits its indexed sites. This is, therefore, a procedural issue rather than a theoretical issue.

From a search engine provider's point of view, a human can determine how relevant a page is to the search keyword, though this may be somewhat subjective. Most search engines nowadays avoid this time consuming process and opt for automated relevancy determination using ranking algorithms. A successful search engine almost relies entirely on how effective the ranking algorithm is [46].

From the perspective of evaluating the relevance of a hit, one can use a rare word as the keyword. Such search usually returns a limited number of hits and therefore it is feasible to examine the relevance of each hit manually. This also makes the measurement of the effectiveness of the ranking algorithm possible by comparing the opinions of the humans and the ranking algorithm.

3.6 Discussions

In this chapter, we introduce our search engine evaluation model, and describe the selected parameters in details. The advantage of this model is that it can be tailored to an individual's needs by changing the weight assignment, deleting parameters, and adding parameters. A user can also focus on the evaluation of a particular aspect easily due to the model's hierarchical structure. We introduce the common list as the benchmark in our model. Currently, we use a simple method to generate the common list, but further investigations on the concept of the common list and the exploration of more sophisticated algorithms are necessary. In next chapter, a case study using this evaluation model is presented.

Chapter 4

Evaluation of Chinese Search Engines: A case study

The rapid development of the Chinese web has made Chinese search engine an indispensable tool for Chinese Internet users. The Chinese language is very different than English and as a result Chinese search engines have their own unique features. Consequently, developing effective Chinese search engines faces many challenges. This chapter presents a case study of our search engine evaluation model: Chinese search engine evaluation. The Chinese web and Chinese search engines are introduced. Previous works on Chinese search engine comparison are described. Evaluations of Chinese search engines using our model are presented. The results obtained from the automated method using a common list are compared to those obtained from human evaluation.

4.1 Overview of the Chinese Web

Over the past few years, Internet development in China has been phenomenal. It is expected that the majority of web pages will be written in Chinese in the very near future. The numbers of Chinese web pages and Chinese Internet users have increased rapidly. In 1997, there were only 300,000 computers connected to the Internet in China. At the beginning of 2004, there were 79.5 million Internet users [61]. By the end of 2004 [62], there were more than 87 million users browsing close to 600,000 web sites and 5.4 million pages, with about 1.2 million domain names [6]. Table 4.1 and Figure 4.1 show the changes in the number of Internet users, computer hosts and "www" websites from 2004 to 2006 [66] [65] [64] [63] [62] [61].

	2004/2	2004/7	2005/1	2005/7	2006/1	2006/7
Internet Users: (million)	79.50	87.00	94.00	103.00	111.00	123.00
Computer Hosts: (million)	30.89	36.30	41.60	45.60	49.50	54.50
"WWW" Websites (including .cn, .com, .net and .irg) (*10,000)	59.56	62.66	66.89	67.75	69.42	78.84

Table 4.1 *Chinese web development according to CNNIC's survey reports*

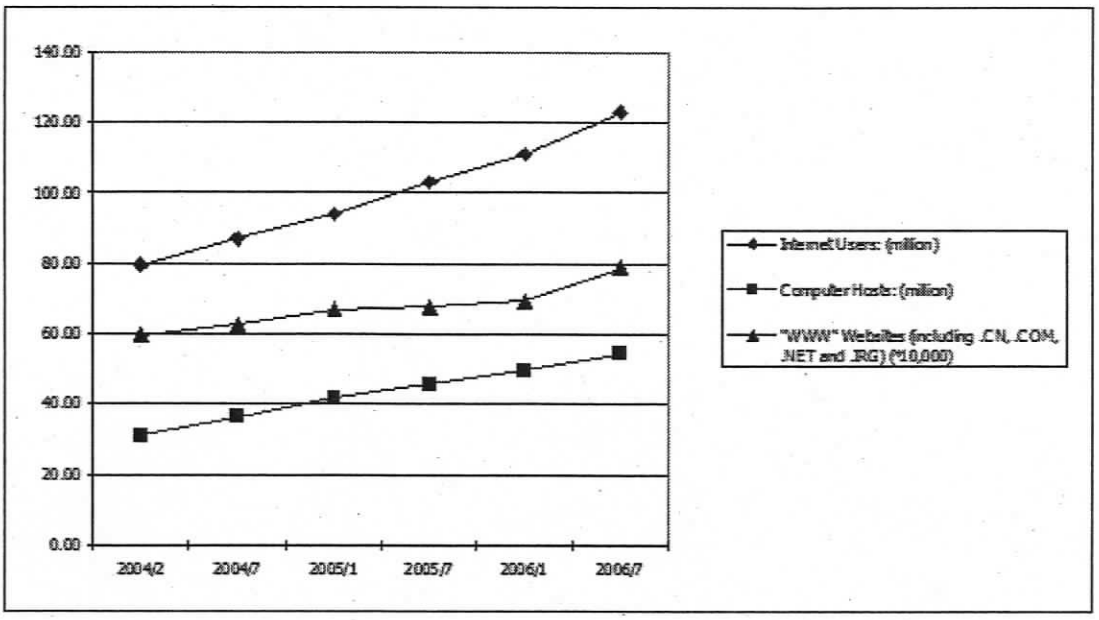


Figure 4.1 *Trend of Chinese web development according to CNNIC's (China Internet Network Information Center) survey reports*

In 2006, China had 123 million Internet users and ranked second in Internet user population. Yet, that was only 9.4% of China's population. This projects a huge potential market for the Chinese Internet [10], and makes Chinese search engine comparison an interesting exercise.

4.2 Overview of Chinese Search Engines

Web mining involves three main areas: content mining, usage mining and web structure mining. Of these three, content mining is the most difficult for the Chinese language as compared to structure and usage mining. This has a direct impact on the quality of the search results. There are several issues that make Chinese web search and document processing much more challenging than those for the English web. Unlike English, there are many different Chinese character sets in use, depending on the geographic region of the web site and the political preference of the author. Big Five (BIG5) or *Dawu*, the traditional Chinese character set, is used in Taiwan and Hong Kong. In China, GB or *Guojia Biaozhun* (National Standard) is used to represent simplified Chinese characters. Increasingly, new web sites either use GBK, *Guojia Biaozhun Kuozhan* (Extended National Standard), or the multilingual Unicode Standard, both of which contain a larger

character set that includes GB and BIG5. Reference [7] gives a comprehensive overview of Chinese character sets and encoding.

The second problem associated with Chinese language processing is the absence of white space between words in contrast to English language. Depending on how one reads a sentence or combines the characters, it is possible to have multiple valid interpretations of the same phrase or sentence. Therefore, Chinese word (or character; bigram, a word that is written with two Chinese character; trigram, a word that is written with three Chinese character, etc.) segmentation is a very difficult problem [16]. The effectiveness of the term extraction process, the clustering of similar documents and the categorization of documents, affect the search engine's capability of indexing its document database in an optimal fashion [28]. In addition, the white space problem also occurs at the keyword entry level that directly affects how accurate the matching will be. Similar issues exist for several other Asian languages such as the Japanese Kanji writing system.

4.3 Previous Works on Chinese Search Engine Comparison

Currently, there are more than 300 active Chinese search engines. It is very important to have a reliable method to compare these search engines to ensure efficient and effective web browsing by different users. However, there are only a few reports in the literature that discuss Chinese search engine comparison and ranking [55].

One of the first publications was [26] in 1997 where Kingoff observed that the reviewed Chinese search engines did not have significant overlaps in the first result page. He concluded that these search engines were different in their search focus, and each had its own niche.

A more thorough investigation was made by the Shanghai Society for Scientific and Technical Information [52]. Twenty-one Chinese search engines were studied and compared based on topic classification, result ranking, hit recentness, page summarization, and coding support. However, no ranking or scoring was given to the compared engines. Though this article was published in 1998, it remains as one of the most complete surveys on Chinese search engines. Another article published in the same year in eSAS World introduced twenty-six Chinese search engines, and described each

engine's features. The author recommended Openfind, TianWang, and Yahoo China [35] without any rigorous comparison.

The Popular Computer Week magazine published a report on five commonly used search engines in June 2000: Yahoo China, Sohu, Goyoyo, Zhaodaole, and Tonghua [42]. Parameters for comparison included home page features, search options, and keyword entry options. Yahoo China and Sohu were the top-ranked engines.

In August 2003, PC Computing published a comparison of ten search engines (Sina, Sohu, Netease, Chinaren, Wander, Excite China, Yahoo China, Cseek, Tianwang, and Zhaodaole) using various parameters including home page feature, advance search feature, coding support, dead links, total hits, search speed, search result's relevance, precision, and ranking [41]. Sina was ranked the best search engine. An email survey to Chinese netizens through iUserSurvey in December 2003 found the top three search engines being Baidu, Google, and 3721 [20].

In a report provided by the Tsinghua IT Usability Lab in July 2004 [54], Google, Yahoo China, Baidu, and Zhongsou were compared based on search result's relevancy, recall, and number of dead links. Baidu and Google excelled in this short and simple comparison.

Most of these works showed qualitative comparisons of different features of the search engines under study, without giving any numerical scoring or ranking. They focused on specific aspects of the search engines and considered only a few evaluation parameters. These non-mathematical approaches quite often introduce biased subjectivity, inconsistency, and non-deterministic results. Consequently, the recommendations from these studies regarding the top search engines were very often differed. A systematic and rigorous approach is necessary to develop a consistent way to evaluate and rank Chinese search engines.

4.4 Methods for Data Collection and Search Engine Evaluation

Chapter 3 presents a search engine evaluation model. As an illustration, five popular Chinese search engines are evaluated and compared using this model. This section explains the search engine selection process, the keyword selection process, and data collection rules.

4.4.1 Selection of Search Engines

There are more than 300 Chinese search engines and sites, and this number is rapidly increasing. A search site not only has a search engine's capability, but also contains other specialized information such as for entertainment and electronic goods. It would be very difficult if not impossible to compare all those search engines (in this work, search engines and search sites are sometimes used interchangeably unless there is a need to specify the exact term.), as it would require a huge amount of time and effort. In order to make the comparison manageable, it is important to carefully select only the most important search engines for illustration purpose.

The most important criterion for the search engine selection is their popularity and acceptance by the users. We targeted our evaluation for well known and popular search engines from Mainland China (simplified Chinese), Taiwan (traditional Chinese), and Hong Kong (simplified and traditional Chinese).

Depending on the operation mechanism, search engines can be classified into three different types: subject directory search engine, full text search engine, and meta-search engine. The most well known examples of these three types are Yahoo China, Google China, and Widewaysearch, respectively. In China, there are many portal websites, for examples, Sina and Sohu, which have an extensive directory structure. In addition, they also support full text search capability utilizing third party support. For example, Google China is a full text search engine. It uses its own technology and very often provides search capability to other web portals. Yahoo China also used Google for searching web pages in the past. Now, however, Yahoo China has developed its own search technology and provides full text search service in addition to subject directory search. Meta-search engines, though not that many, also exist in the Chinese web. Meta-search engines do not crawl the web to compile their own searchable database. Instead, they search the databases of the other search engines simultaneously, and provide the top ranked results from each search engine to the users. In this work, search engines belonging to each of these three types were reviewed.

Besides the general-purpose search engines, there are also many specialty search engines. For example, Cha Jia (http://www.chajia.com/) focuses on searching the price of

electrical products. This research concentrates on general-purpose search engines and therefore specialty search engines are not included in our study.

We surveyed the most commonly used search sites and sites that had many attractive search features. After extensive browsing and searching on the web, as well as reviewing the many topic directories, we identified forty-two prominent Chinese search engines. Table 4.2 lists these search engines along with their URL, the host engine used, and their geographical regions. A '*' is found in the cell under the 'Engine used' column if the search engine used could not be identified for that search site. Of these 42 engines, 18 have the affix CN (China), 11 have the affix HK (Hong Kong), and 13 has TW (Taiwan).

Site	URL	Engine used	Region
Google China	http://www.google.cn	Google	CN
Yahoo China	http://www.yahoo.com.cn/	Yahoo	CN
Bai Du	http://www.baidu.com.cn/	Bai Du	CN
ZhongGuoSouSuo	http://www.zhongsou.com/	ZhongGuoSouSuo	CN
BeiDaTianWang	http://e.pku.edu.cn/	Tian Wang	CN
Sina	http://cha.sina.com.cn/	ZhongGuoSouSuo	CN
Sohu	http://dir.sohu.com/	ZhongGuoSouSuo	CN
NetEase	http://search.163.com/	ZhongGuoSouSuo	CN
Altavista China	http://www.altavista.com/	Altavista	CN
Goyoyo	http://www.goyoyo.com/	BaiDu	CN
BeiJiXing	http://www.beijixing.com.cn/	Beijixing	CN
21CN.COM	http://cha.21cn.com/index.html	BaiDu	CN
Widewaysearch	http://www.widewaysearch.com	Meta-search engine	CN
Shalala	http://www.shalala.net/	*	CN
Lycos Asia	http://cn.lycosasia.com/	Baidu	CN
China.com	http://www.china.com/	Yahoo	CN
MSN CN	http://china.msn.com/	3721, 3721 has been bought by YAHOO	CN
Net2asp	http://www.net2asp.com.cn/	5414 (its own technology)	CN
Google HK	http://www.google.com.hk/	Google	HK

Table 4.2 A list of popular general-purpose Chinese search engines (continue on next page)

Site	URL	Engine used	Region
Yahoo HK	http://hksar.hki.yahoo.com/	Yahoo (YST)	HK
Openfind	http://cd.openfind.com.tw/HK-CD/	Openfind	HK
Lycos Asia HK	http://www.myrice.com/	Lycos	HK
MSN HK	http://search.msn.com.hk/	MSN	HK
Sina HK	http://cha.sina.com.hk/	Google	HK
WebinHK	http://www.webinhk.com/	*	HK
GreenWorld	http://ep.sunup.net/index.php	*	HK
TimWay	http://www.timway.com/	Openfind	HK
36it	http://36it.com/search/pages/	Meta-search engine	HK
Shalala	http://www.shalala.net/	Google	HK
Google TW	http://www.google.com.tw/	Google	TW
Yahoo Kimo	http://tw.yahoo.com/	Yahoo (YST)	TW
Openfind	http://www.openfind.com/taiwan/index.php	Openfind	TW
Sina TW	http://www.sina.com.tw/	Google	TW
Yam.com	http://www.yam.com/	Google	TW
Wasite.com	http://www.wasite.com/	dictionary search	TW
GreenWorld	http://ep.sunup.net/index.php	*	TW
Pchome	http://dir.pchome.com.tw/	Openfind	TW
Gais	http://gais.cs.ccu.edu.tw/	Gais	TW
Formosa	http://www.formosa.com.tw/	*	TW
Don-Net	http://www.don-net.com.tw/	dictionary search	TW
Nobel	http://www.nobel.com.tw/	*	TW
Seeder	http://www.seeder.net/	Use PChome and Google results	TW

Table 4.2 *A list of popular general-purpose Chinese search engines*

The list of forty-two search sites was still too large for a thorough and meaningful investigation. Some search engines were further eliminated according to the following criteria:

- Many search engines uses another search engine as the host. For example, popular sites such as Shalala and Yam use Google's search engine. All the engines that do not have their own search mechanism were eliminated.

- Sites that are not readily accessible, either because they are too busy (slow) or the time to download the home page is relatively long, were eliminated from the comparison.
- Many search sites in Taiwan and Hong Kong are poorly designed and unusable due to their long response time and limited capability. Furthermore, most of the short-listed sites in these two regions are powered by their equivalent in Mainland China. For instance, a search on Google.HK and Google.CN yield almost the same results. Therefore, we decided to focus on the evaluation on search engines from Mainland China.

This process of elimination left five search engines for evaluation. A brief description of each of the selected engines is given below (statistics cited as of June 21, 2007):

- **Google China** (http://www.google.com/intl/zh-CN/): established on September 12, 2000, it has over 6 billion home pages, 880 million images, 845 million news, and many search options and features.
- **Yahoo Yisou** (http://www.yisou.com/): with the version released on June 21, 2004, it has over 8 billion web pages, 550 million images, 1 million music pieces, articles available in 38 languages, and many search options and features.
- **Zong guo sou suo** (http://www.zhongsou.com): founded in September 2002 as the first Chinese search engine that supports trade classification, it claims to have 2.8 billion web pages and supports homonym rectification.
- **Baidu** (http://www.baidu.com): founded in 1999, it claims to have more than 8 billion web pages, over 600,000 MP3 and more than 500 news sites; it provides services such as algorithmic search, enterprise search and pay-for-performance; it has a large search range including Hong Kong, Taiwan, Macao, Singapore, and some web sites in North America and Europe; it also supports homonym rectification.
- **Tianwang** (http://e.pku.edu.cn): founded in October, 1997, it is very popular among academics; this Peking University site has over 6 billion home pages and can search other Chinese university sites as well as over 1,000 American university sites; it also supports ftp search.

4.4.2 Selection of Chinese Keywords

In order to make the analysis manageable and meaningful, rare words are often used for search engine evaluation. This limits the number of hits and makes it easier to test the capability and effectiveness of the search engines. For examples, 'crumpet' and 'polyphenol' were used in [38], and ten rare words were used in Ljosland's study [27]. Strong query (query that is designed to strongly identify the page) [36] is often used to ensure the page containing the match would appear in the search results.

Two phrases and three of their variations were selected as keywords in this study as shown in Table 4.3. The first one is 'Chinese Search Engine Comparison' (b), an appropriate choice within the context of this work. The second phrase is 'Bird Flu' (e), a hot news item since the beginning of 2004. The variations are used to 'fool' the search engines into finding mismatched items, as a preliminary test on the segmentation capability of these search engines.

(a)	中文搜索引擎比较	A free search on 'Chinese search engine comparison'
(b)	"中文搜索引擎比较"	An exact search of 'Chinese search engine comparison'
(c)	"中文搜索引擎" "比较"	Search using two exact phrases of 'Chinese search engine' and 'comparison'
(d)	禽流感	A free search on 'bird flu'
(e)	"禽流感"	An exact search on 'bird flu'

Table 4.3 *Keywords used in this case study*

4.4.3 Data Collection Methodology

There is no unanimous opinion on the methodology of data collection for search engine comparison. Many researchers argue that data should be collected during off-peak hours [22]. During peak hours, search engines receive a lot of queries at the same time and the busy search engines may become slower or even time out before returning the results. We disagree with this argument. To be attractive to users, a search engine should have relatively stable performance regardless of the demand. We believe that data collected

over different periods of time, including both peak hours and off-peak hours, would give a better indication of the efficiency and effectiveness of a search engine.

To compare and evaluate the five search engines, data were collected over two weeks, from August 8, 2004 to August 21, 2004. During weekends, data were collected on Sundays; during weekdays, data were collected on Wednesdays. For each day, three samples were collected at Beijing time 10:00 AM, 9:00 PM, and 1:00 AM. The reasons for selecting these three time instances were to avoid the bias for peak time and off-peak time collection, and also to avoid the bias for day time and night time samplings. For each sample, three rounds of search were performed using the five keywords in each of the five selected search engines. Each round was separated by 30 minutes and only the first result pages were used for evaluation. The total number of data sets was 3(rounds) X 3(times) X 2(days) X 2(weeks) giving 36 sets, and the total number of searches was 5(keywords) X 5(websites) X 3(rounds) X 12(times) resulting in 900 pages.

4.4.4 Methods of Evaluation

Our search engine evaluation model measures two groups of parameters: feature and performance. The parameters in the feature group are qualitative. It is relatively straightforward to obtain the score for the feature group. Evaluating score for the performance is more complicated. The performance criteria selected for the comparison are relevance, number of hits, response time and number of problems. Relevance is subjective and needs to be rated by humans. For this purpose, three volunteers were solicited. The individuals were undergraduate Chinese students at the University of Victoria. They are comfortable with using search engines but are not experts doing research on search engines. They can be considered as typical users of Chinese search engines. The search results from the five search engines were given to them and they were asked to rate the search engines according to the relevance of the returned results. No guideline was given as what relevance is in order to maintain the individual's subjectivity. We requested the volunteers to assign a score from 1 to 10. All volunteers' results for each keyword were averaged and normalized for comparison, and used for the 'quality of result' parameter in Figures 4.4, 4.5 and 4.8.

Instead of using human judgement for the relevance parameter, one can use the concept of a common list to derive precision and recall scores as described in Section 3.4.2. There are two ways to make up the common list. One way is to use the opinions of the volunteers to form a "Human list". A common "Algorithmic list" can also be obtained using an algorithmic approach. The procedures to generate these lists are described in Section 4.5.5.

4.5 Analysis of Results

Through the data analysis process, we expect to discover average patterns and the degree of variations on response time, as well as update frequency as indicated in any changes in the number of hits in these thirty-six sets of data. The results show that Google and Yahoo updated most frequently, while Tianwang did not update its database throughout our experimentation period.

4.5.1 Chinese Language Specific Issues

The results from all five search engines exhibit the peculiarity of the Chinese language. As expected, results of (b) from Table 4.3 were limited while (a) returned items including the ones in (b) and (c). In addition, (a) also returned items with the independent phrases of 'Chinese', 'search', and 'engine comparison', in which engine was interpreted in the machinery sense. All five search engines exhibit this behaviour. Similarly in (d), retrieved documents had 'bird' (the first character) and 'flu' (the second and third character together) appeared together and also separately. This rendered the results in two categories either of which may suit the need of the particular user. This pattern, however, is valid for all languages to some extent.

To further examine how each engine handles Chinese phrases, we performed further experiments using additional keywords as shown in Table 4.4. AIDS is a very specialized word in Chinese and as expected, the free and exact search of AIDS (i) produced similar results. The left-hand-side word in (ii) is used for 'angel' in traditional literature, while the word on the right hand side is its modern English phonetic representation. Both forms are widely used these days. The results, as expected, indexed the two variations into two distinct categories of documents. Similarly, the two variations

of ice cream (iii), both are English phonetic representations, resulted in two groups of non-overlapped documents. Finally, in (iv), two sets of documents each referencing a famous author either by his real name or pen name were returned, with some overlaps. These results point to the need of a Chinese synonym database to make the search more effective.

(i)	艾滋病 / "艾滋病"	A free search and an exact search of AIDS
(ii)	天使 / 安琪儿	The two variations of angel
(iii)	冰激凌 / 冰淇淋	The two variations of ice cream
(iv)	鲁迅 / 周树人	The pen name and real name of a famous author

Table 4.4 *Additional keywords*

4.5.2. Response Time and Number of Results

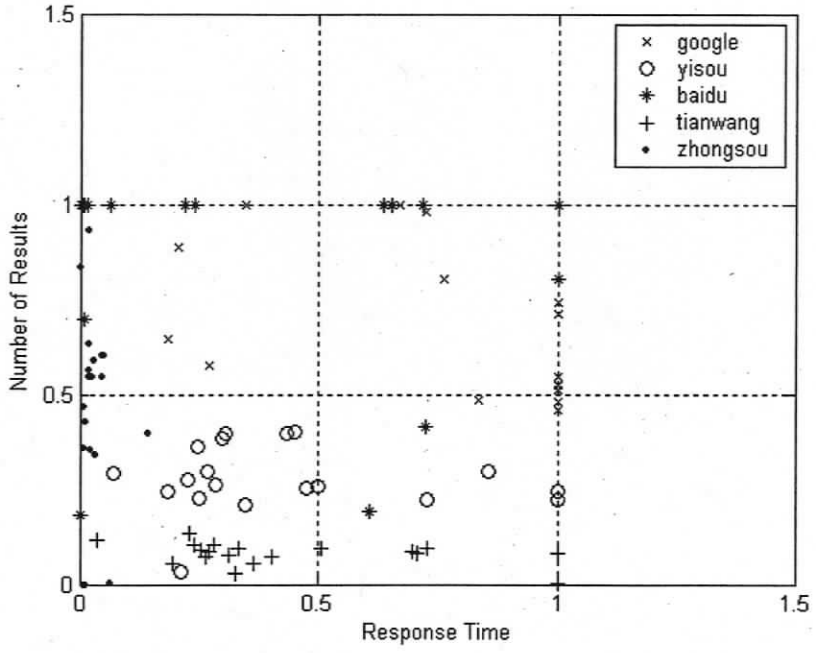


Figure 4.2 *Response time versus number of results*

Figure 4.2 shows a scatter plot of the response time versus the number of results for all the keywords used. For each keyword, its response time and number of results are normalized to the longest response time and the largest number of results among all keywords' results. Ideally a good search engine should locate the largest number of

documents in the shortest time. This corresponds to the upper-left quadrant in the figure. For this quantitative measure, Zhongsou and Baidu performed the best.

4.5.3 Features Comparison

In Chapter 3, the hierarchy of the feature parameters group is described. Tables 4.5, 4.6, 4.7, 4.8, 4.9, and 4.10 show the values of the feature parameters for the five selected engines.

Description		Possible values	Sign of Weight	Results				
				Google	Yahoo	Baidu	Zhong sou	Tian wang
Homepage features								
	User's evaluation = number of average (U1+U2+U3)	normalized number	"+"	1	1	0.84	0.73	0.77
	Help link	yes=1, no=0	"+"	1	1	1	1	1
	Result language selection	yes=1, no=0	"+"	1	1	0	0	0
	Directory search selection	yes=1, no=0	"+"	1	1	1	1	1
	Advanced search selection	yes=1, no=0	"+"	1	1	1	1	0

Table 4.5 *Values of the feature parameters – homepage features*

Description		Possible values	Sign of Weight	Results				
				Google	Yahoo	Baidu	Zhong sou	Tian wang
Database								
	Directory: normalized number of categories	normalized number	"+"	0.941	0.82	0	1	0.65
	Database: normalized total number of pages	normalized number	"+"	0.86	1	0.06	0.04	0.02

Table 4.6 *Values of the feature parameters – database*

Description		Possible values	Sign of Weight	Results				
				Google	Yahoo	Baidu	Zhong sou	Tian wang
User preferences								
	Homepage Interface Language (display the tips and message in which languages)	yes=1, no=0	"+"	1	0	0	0	0
	Safe Search Filtering	yes=1, no=0	"+"	1	1	0	0	0
	Number of Results (number of results per page)	yes=1, no=0	"+"	1	1	1	1	0
	Results Window (in the new window)	yes=1, no=0	"+"	1	1	1	1	0
	Intelligent input correction	yes=1, no=0	"+"	0	0	0	1	0
	Set the search default homepage	yes=1, no=0	"+"	0	0	0	1	0
	The Search Option in the result's page	yes=1, no=0	"+"	0	0	0	1	0
	News display capability in search result page	yes=1, no=0	"+"	0	0	1	0	0

Table 4.7 *Values of the feature parameters – user preferences*

Description			Possible values	Sign of Weight	Results				
					Google	Yahoo	Baidu	Zhong sou	Tian wang
Keyword entry options									
	Stop word		yes=1, no=0	"+"	1	1	1	1	1
	Case sensitivity		yes=1, no=0	"-"	0	0	0	0	0
	Exact phase (use quotation mark)		yes=1, no=0	"+"	1	1	1	1	1
	Asterisk wildcard (e.g. "*", "?")		yes=1, no=0	"+"	1	1	1	0	0
	Search by pronunciation		yes=1, no=0	"+"	1	1	1	1	0
	Boolean operators								
		AND/and/"+"/" "	yes=1, no=0	"+"	1	1	1	1	1
		OR/or/"|"/"/"	yes=1, no=0	"+"	1	0	1	1	0
		NOT/not/"-"	yes=1, no=0	"+"	1	1	1	1	0
	Others		yes=1, no=0	"+"	0	0	0	0	0

Table 4.8 *Values of the feature parameters – keyword entry options*

Description		Possible values	Sign of Weight	Results				
				Google	Yahoo	Baidu	Zhong sou	Tian wang
Results								
	Total hits indication	yes=1, no=0	"+"	1	1	1	1	1
	Number of pages indication	yes=1, no=0	"+"	1	1	1	1	1
	Show search time	yes=1, no=0	"+"	1	1	1	1	1
	Search within the results	yes=1, no=0	"+"	1	1	1	1	1
	Results are ordered	yes=1, no=0	"+"	1	1	1	1	1
	Results are numbered	yes=1, no=0	"+"	0	0	0	0	1
	Results in various file formats	yes=1, no=0	"+"	1	0	0	0	1
	Paid listing/result	yes=1, no=0	"-"	0	1	1	1	0
	Web page snap	yes=1, no=0	"+"	1	1	1	1	1
	Related pages	yes=1, no=0	"+"	1	0	0	1	0
	Check pages in the same hit's website	yes=1, no=0	"+"	0	1	1	1	0
	Date of the hit	yes=1, no=0	"+"	0	1	1	0	1
	Size of the hit	yes=1, no=0	"+"	1	1	1	0	1
	Comments of the hits	yes=1, no=0	"+"	0	0	0	0	0

Table 4.9 *Values of the feature parameters – results*

Description		Possible values	Sign of Weight	Results				
				Google	Yahoo	Baidu	Zhong sou	Tian wang
Search options								
Search modifier								
	Include all of the following keywords	yes=1, no=0	"+"	1	0	1	1	0
	Include the following exact phrase	yes=1, no=0	"+"	1	0	1	0	0
	Exclude	yes=1, no=0	"+"	1	0	1	1	0
	At least one of	yes=1, no=0	"+"	1	0	1	1	0
	Case sensitivity specification	yes=1, no=0	"+"	0	0	0	0	0
	Others	yes=1, no=0	"+"	0	0	0	0	0
Search field (return results where the terms occur)								
	Anywhere in the page	yes=1, no=0	"+"	1	1	1	1	1
	In the title of the page	yes=1, no=0	"+"	1	0	1	0	0
	In the text of the page	yes=1, no=0	"+"	1	0	0	0	0
	In URL of the page	yes=1, no=0	"+"	1	0	1	0	0
	In the links of the page	yes=1, no=0	"+"	1	0	0	0	0
	Others	yes=1, no=0	"+"	0	0	0	0	0
Search focus selection								
	Website	yes=1, no=0	"+"	1	1	1	1	1
	Web page	yes=1, no=0	"+"	1	1	1	1	1
	Directory	yes=1, no=0	"+"	1	1	1	1	1
	Others (e.g. mp3, news, .)	yes=1, no=0	"+"	1	1	1	1	1

Table 4.10 Values of the feature parameters – search options (continue on next page)

Description		Possible values	Sign of Weight	Results				
				Google	Yahoo	Baidu	Zhong sou	Tian wang
Search constraint selection								
	Result language selection	yes=1, no=0	"+"	1	1	1	0	0
	Result file format	yes=1, no=0	"+"	1	1	0	0	1
	Time limiting capability	yes=1, no=0	"+"	1	1	1	0	0
Search meta words								
	Site: return results from the specified URL or domain	yes=1, no=0	"+"	1	1	1	1	0
	Similar: return pages that are similar to the hit	yes=1, no=0	"+"	1	0	0	0	0
	Links: find pages that link to the hit	yes=1, no=0	"+"	1	0	0	1	0
	Geographic region: search within the selected ranges	yes=1, no=0	"+"	0	0	1	0	0
	Others (e.g.inurl, ntitle,…..)	yes=1, no=0	"+"	1	0	1	0	0

Table 4.10 *Values of the feature parameters – search options*

Table 4.11 and Figure 4.3 present and compare the results of each major feature group for the five selected search engine. For illustration purpose, we assigned different weights to the six groups as shown in Table 4.11. Following the order listed, the weights assigned are 0.1, 0.1, 0.3, 0.2, 0.1, and 0.2, respectively. The flexibility of having different search options is considered the most important factor among the six groups. Table 4.11 tabulates the results of the features group. Figure 4.3 shows the corresponding histograms for each feature group. It can be seen that different search engine excels in different search features.

		Weight	Parameter value				
			Google	Yahoo	Baidu	Zhong sou	Tian wang
Features group score			**0.8**	**0.6**	**0.6**	**0.5**	**0.4**
	Home page features	0.1	1.0	1.0	0.8	0.7	0.6
	User preferences	0.1	0.5	0.4	0.4	0.6	0.0
	Search options	0.3	0.9	0.5	0.7	0.4	0.3
	Keyword entry options	0. 2	0.7	0.7	0.7	0.6	0.3
	Database	0. 1	0. 9	0. 9	0. 0	0. 5	0. 3
	Result features	0. 2	0.6	0.6	0.6	0.5	0.7

Table 4.11 *Feature group score*

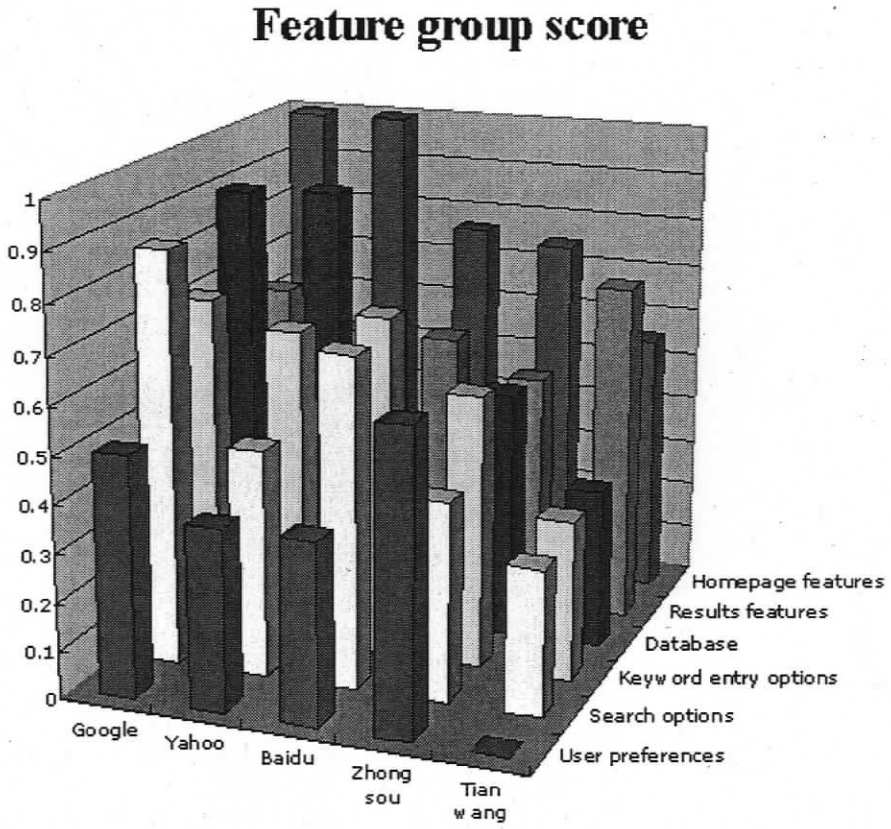


Figure 4.3 *Feature group score*

4.5.4. Performance and Overall Comparison Based on Human Evaluation

As mentioned in Section 4.4.4, volunteers were asked to rate the relevance of the collected data. Evaluation of the other parameters (problems. response time, number of

hits) is not subjective and they were obtained easily from the collected data.

Performance comparison

Figure 4.4 shows the scores of the components in the performance group. For each parameter, the scores from different engines are normalized with respect to the highest score. Tianwang performs very poorly in response time and number of hits. This is not very surprising since this engine is intended for a special group, i.e., academics. Zhongsou excels in response time but ranks very poorly in human-evaluated relevance. Google and Yahoo have mediocre to good performance in all the categories. Knowing these characteristics, a user can choose the appropriate search engine to suit his/her needs.

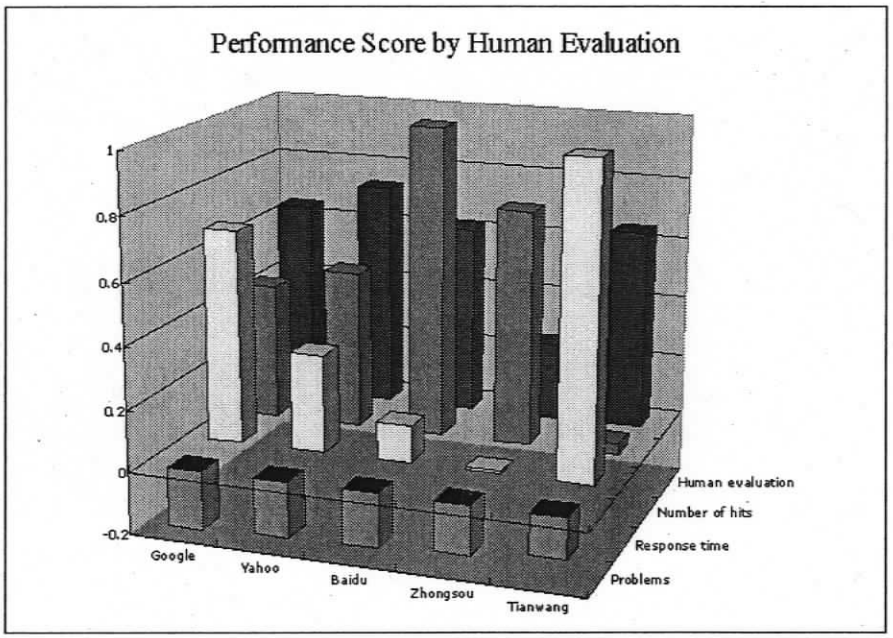


Figure 4.4 *Performance with human evaluation*

Overall comparison

Figure 4.5 shows a comparison of the overall score with human-evaluated quality of results (i.e., relevance). Google has the highest score as it is evident that it has many good features as compared to the other search engines.

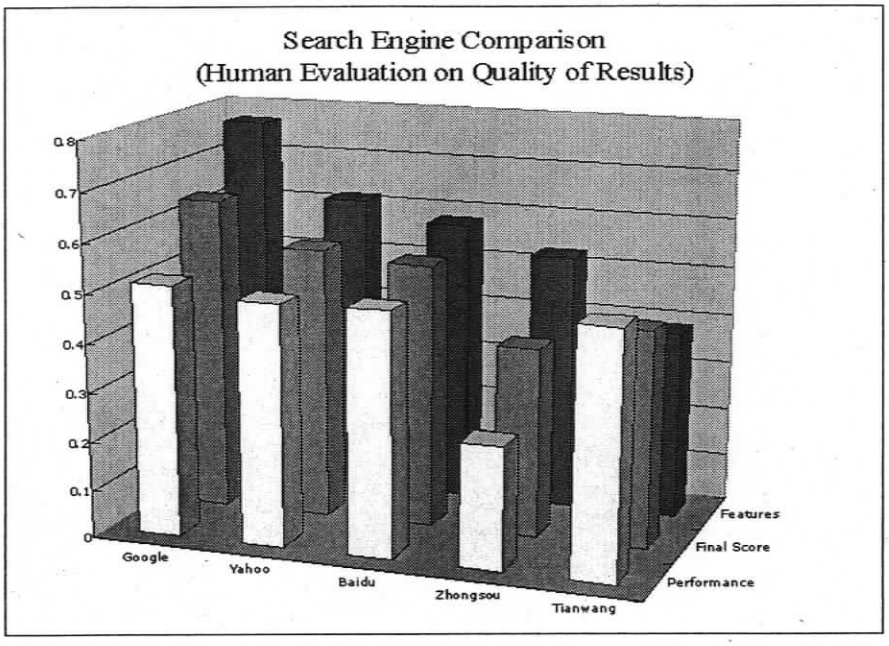


Figure 4.5 *Overall comparison of the search engines*

4.5.5. Search Engine Comparison Based on the Common List

In Chapter 3, the concept of a common list is introduced. Precision and recall are derived from the common list. To generate the human list, our three volunteers were asked to create a 25-item list according to their own judgement using the results from the five search engines, which they thought are relevant to the keyword.

In this research, a simple algorithm is used to compose the common algorithmic list. To maintain a size of roughly 25 (arbitrarily decided) items in the algorithmic list, the five top-ranked items from each engine are used. If the total number of items in this initial attempt is less than 25 due to overlaps, then the top 6 items from each search engine are considered, and then the top 7 items, and so on. This iterative process concludes when the common list has 25 or more relevant items. In essence, this common list is very similar to the recommended list generated by a meta-search engine. Results from our simple algorithm reported here are for exploratory and illustration purposes. In real life situations, any of the more sophisticated existing algorithms, such as the ones in [51] and [30], can be utilized to merge the individual ordered lists into a common ranked

list. A second common list, the human list, is obtained in a similar fashion using the individual lists (with 25 items each) as drawn up by the volunteers.

Figure 4.6 shows the relevance rating as discussed based on a human list, while Figure 4.7 shows the relevance rating based on an algorithmic list. From the histograms, we can see the results are similar in using these two common lists. This initial and preliminary experiment shows that an algorithmic list, which can be automated, is comparable to a human list.

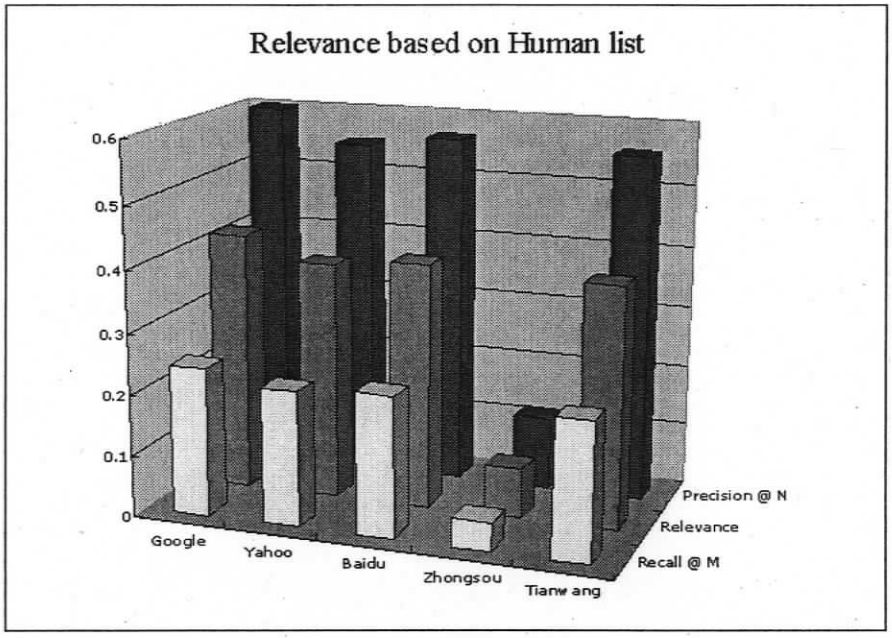


Figure 4.6 *Relevance rating based on a human list*

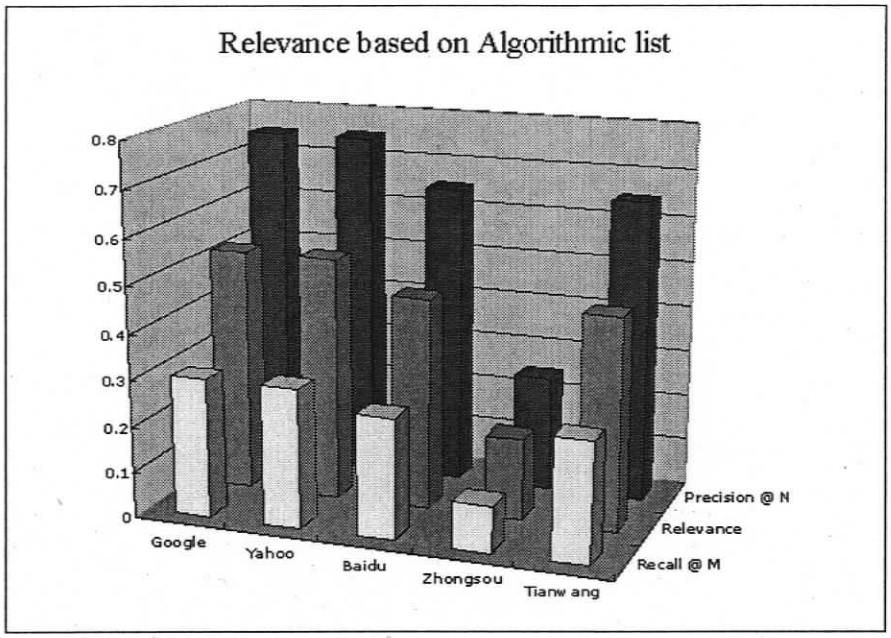


Figure 4.7 *Relevance rating based on an algorithmic list*

4.5.6 Comparison of Results Obtained from Different Methods

Finally, we would like to examine whether there is significant discrepancy among the relevance factors obtained using the three different methods. Figure 4.8 shows relevance comparison among the five search engines using different evaluation methods. Though the magnitudes are different, the rankings of the five search engines are similar in all three methods. This gives us confidence on the validity of the common list approach. It is worthwhile to pursue further the ideas of generating a meaningful common list and automating the evaluation process.

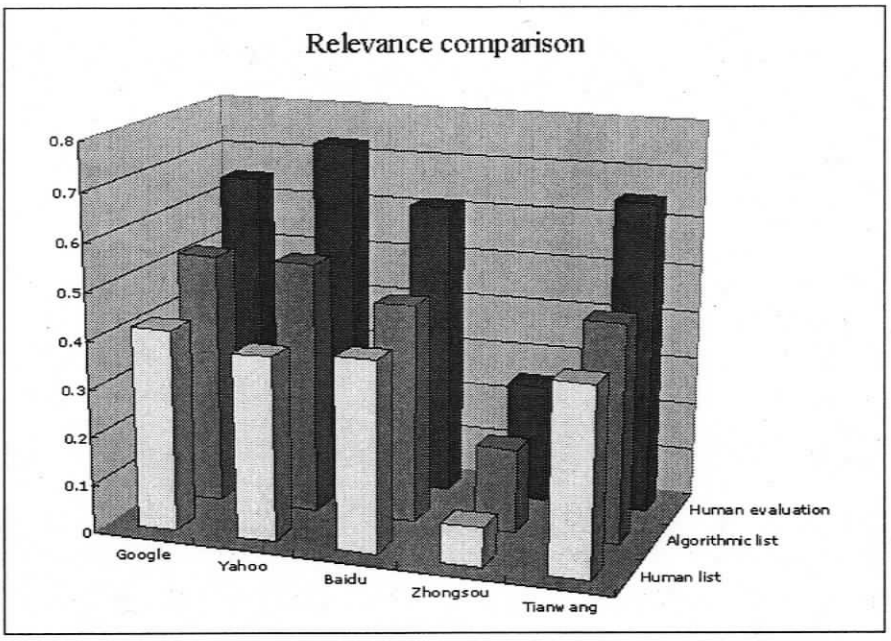


Figure 4.8 *Relevance comparison of the search engines using different evaluation methods*

4.6 Conclusions

This chapter presents a case study of the proposed search engine evaluation model. Five Chinese search engines are evaluated and ranked. The results show the methodology is valid and the development of an automated, non-subjective search engine evaluation process deserves further investigation.

Chapter 5

Analysis and Comparison of Search Engine Performance over Time

The web is very dynamic and is changing every moment. The performance of a search engine depends on many factors, the major ones being the rate at which the database is updated and the effectiveness of the ranking algorithm. These aspects have not been investigated in details except for a few relatively recent reports by [22] [25] [13]. The aim of this chapter is to investigate methodologies to quantitatively compare results obtained from different search engines over a survey period of time.

Section 1 briefly reviews existing work in the literature on analyzing and comparing search engine performance over time. Data were collected for both Chinese and English search engines. The data collection process is summarized in Section 2. The three metrics used to analyze the data are presented in Section 3. Histograms are employed to present obtained results using these metrics. Section 4 shows how these histograms are classified into different groups according to their patterns, and speculates the possible reasons for these different patterns with respect to the database update frequency and the ranking algorithm of the search engines. Presenting data using histograms allows visual comparisons, though interpreting and inspecting a large number of histograms is time consuming and difficult. Section 5 introduces a novel method to quantitatively compare different histograms within the context of search engine performance. The merits of the proposed method and its potential applications are discussed in Section 6.

5.1 Previous Work and Motivation

Search engine evaluations can be broadly classified into two groups: qualitative and quantitative. Most studies on search engine evaluation are based on human-based evaluation method [11], [59] and therefore are qualitative in nature. Data collected over a

time period are presented as histograms to allow visual inspection and analysis. Recently, quantitative evaluations of search engines have attracted a lot of interest. These methods compute one or more numbers that reflect the performance of the search engines under examination [25].

The most straight forward measure to compare the variation of results over time of a search engine is to compute the number of overlaps among the top ten items in different days. However, this simple measure fails to specify when these top-ranked items are the same but different in their rankings in different days.

Spearman's footrule [43] [5] is introduced to alleviate this problem. In this approach the items that are not common in the two lists are excluded and two new ranked lists are formed which consist of the common items. The two new lists can be represented as two vectors of equal dimensions. The "distance" between the two vectors are then computed [25] [24] for performance comparison purpose. The limitation of this approach is that it compares only the items common in the two lists and completely disregards the ranks of the items not present in both of the lists.

Fagin et al. proposed a method to correct this problem. In this approach, when comparing two lists of length n, any item presents in one list but not in the other one is assumed to occupy position $n+1$ in the latter list. The problem is that the items that are not common in the two lists have a considerable impact on the performance measurement in Fagin's approach [44].

It is clear from the above discussion that a single measure cannot adequately describe the performance of a search engine. One needs to compute a number of measures to have a better assessment of the performance. This approach was followed by Bar-Ilan and coworkers [25]. They used a combination of overlaps, Spearman's footrule, Fagin's method and a few other methods to analyze the collected data.

In this work we take a different approach to analyze search engine performance. We first derive the update or duplication frequency, and the rank change frequency from the collected data. We use histograms as a visualization tool to quickly get an idea about the characteristics and performance of a search engine with respect to these frequencies. Histograms have been used in previous search engine research work [11], but we use it in a more systematic way. We classify the histograms into different groups according to

their shape, to gain a better understanding of the search engines' behaviour. We also devise a deterministic, quantitative, and automated method for this classification.

5.2 Summary of Data Collection

To investigate the performance of search engines over time, data were collected for a number of Chinese and English search engines using various keywords over several time periods.

The research work presented in this thesis started with the aim of evaluating Chinese search engines and data were collected only for them at the beginning of the project (2005). Later, the scope of the research was extended for other search engines. Data were then collected for both English and Chinese search engines. Tables 5.1 and 5.2 give the summaries of data collection for Chinese and English search engines, respectively. Chao Nv and Furong Sister were selected only for Chinese search engine, since they were hot topics in Chinese media in 2005 and 2006.

Name of the Engines	Keywords	Data collection period
Google, Yisou, Baidu, Tianwang, Zhongsou	Chao Nv, Furong Sister, Hurricane, Katrina, New Orleanbs, Oil Price, Rita, Tsunami, Bird Flu, Hu Jintao, Chen Shuibian, WTO	2005-9-20 to 2005-10-25 2006-9-12 to 2006-11-13

Table 5.1 *Summary of data collection for the Chinese search engines*

Name of the Engines	Keywords	Data collection period
Google, Yahoo, MSN, Lycos, Hotbot, AOL	Hurricane, Katrina, New Orleanbs, Oil Price, Rita, Tsunami, Bird Flu, Hu Jintao, Chen Shuibian, WTO	2006-9-12 to 2006-11-13

Table 5.2 *Summary of data collection for the English search engines*

5.3 Performance Measurement Metrics

There are many different parameters used for evaluating a search engine over time, database updates and rank changes are the most commonly used ones. In this chapter, we focus on these parameters and derive the following three performance measurement

metrics from the collected data:

- Daily duplication frequency: Quantifies items common between two consecutive days.
- Period duplication frequency: Quantifies the occurrence of different items over a period of time.
- Daily rank change frequency: Quantifies the rank change of different items in two consecutive days.

5.3.1 Daily Duplication Frequency

For each search engine, the top 10 results for every two consecutive days (days x and x+1) are compared to determine the number of common items. The normalized duplication number (D_i) is obtained according to the following equation:

$$D_i = \frac{n_i}{T},\ i=1,\ .....,\ 10 \qquad (5.1)$$

Here n_i is the number of times that i items are common between the top ten items in two consecutive days. T is the data collection period. Histograms are then used to visualize this overlap, as shown in Figure 5.1. The height of the bars, i.e., D_i, corresponds to the degree of overlap in consecutive days. The histogram in Figure 5.1 is based on the results obtained from Yisou for the keyword Tsunami. The X-axis shows the possible number of common items, 0 to 10, from day x to day x+1 over T days, and the Y-axis represents the frequency of i items common in two consecutive days. The height of bar 9 is 26.09. This means that during the data collection period, 26.09% of the time, 9 items were common among the top ten items in two consecutive days.

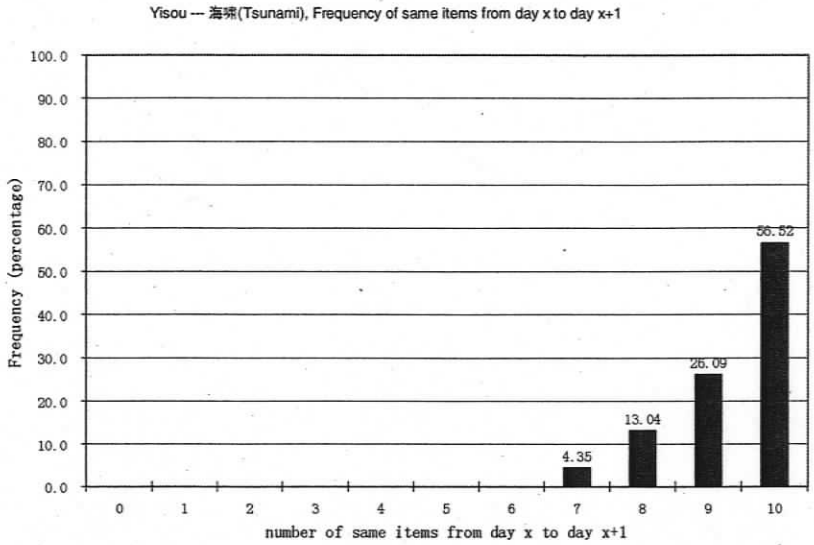


Figure 5.4 *Histograms for daily duplication frequency: Yisou-Tsunami*

5.3.2 Period Duplication Frequency

To obtain useful statistics for this metric, the number of unique items appearing in the data collection period (T) was identified first. These items can appear several times in T days ranging from 1 to T. Q_i is defined as:

$$Q_i = \frac{M_i}{N},\ i = 1, \ldots, T \tag{5.2}$$

Here, M_i is the number of unique items appearing i times during the data collection period in the top ten positions. N is the total number of items ($N = 10 * T$) that appear in the top ten positions during the time period T. A sample histogram is shown in Figure 5.2 for Google using the keyword Chao Nv. The X-axis shows the frequency of occurrence during the data collection period and the Y-axis represents the number of items in percentage. The height of bar 3 is 17.31. This means that 17.31% of the N items appeared 3 times in the top ten positions over the period T.

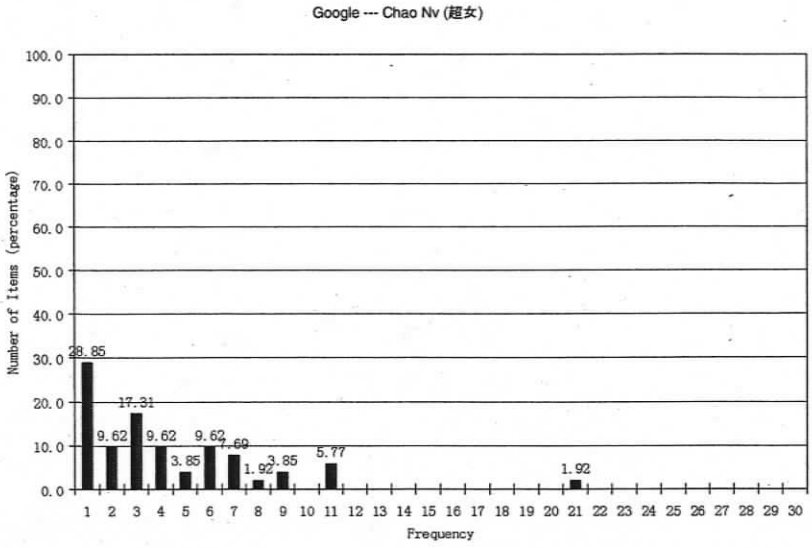


Figure 5.5 *Histograms for period duplication frequency: Google- Chao Nv*

5.3.3 Daily Rank Change Frequency

A key performance criterion for a search engine is how it ranks the relevant pages. To investigate this, results from each day (day x) are compared to those of the next day (day x+1). If item y is in position m in day x and in position k in day x+1, then the change in rank is (m-k). If an item in day x is not on the list of day x+1, the rank change is marked as #, which also indicates that there is a new item on the list. If n items for consecutive days are compared, the n rank-change comparisons (with n*n item comparisons) result in a vector of size $2n$ by 1. The $2n$ indicates either a ranking drop in the range of -1 to -(n-1), no change, a ranking rise in the range of +1 to +(n-1), or a disappearance from the top n. After finishing the comparison over the data set for a period T, T-1 such vectors are available. From those T-1 vectors, the number of times M_i, for each rank change i, from -(n-1) to +(n-1) and #, can be derived. The frequency R_i of the daily rank change is expressed as:

$$R_i = \frac{M_i}{N},\ i = -(n\text{-}1) \ldots +(n\text{-}1), \# \tag{5.3}$$

Here, N is the total number of rank-change comparisons ($N = (T\text{-}1)*n$). In this work, we compare the top ten items between consecutive days. Figure 5.3 shows the daily rank change frequency for search engine Yisou using the keyword Tsunami. The X-axis shows

the rank change from -9 to 9 and #. The Y-axis represents the frequency of the corresponding rank change on consecutive days. The height of Bar 0 is 75.65 which means that there was no rank change among 75.65% out of the *N* rank-change comparisons.

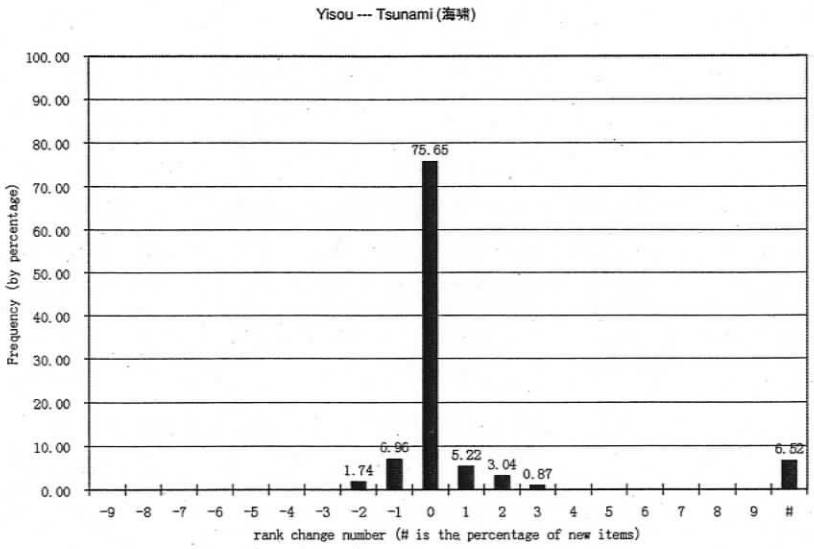


Figure 5.6 *Histograms for daily rank change frequency: Yisou- Tsunami*

5.4 Analysis of Results

The shape or pattern of the histograms provides important information about the search engine's characteristics. After examining the shape of all the histograms, they were classified into different groups. The data collected were rather large in size and the number of histograms obtained was approximately one thousand. Therefore, only representative histograms are presented in the following discussions. The complete set of histograms can be found in Appendix and at http://www.ece.uvic.ca/~wangyl/.

5.4.1 Daily Duplication Frequency

The daily duplication frequency histograms can be classified into three groups according to their shape.

Group DD-R: The bars at the right end have comparatively higher height. An example of this is shown in Figure 5.4. This is the duplication frequency for the keyword Tsunami

from search engine Baidu. This indicates the duplication in two consecutive days is very high.

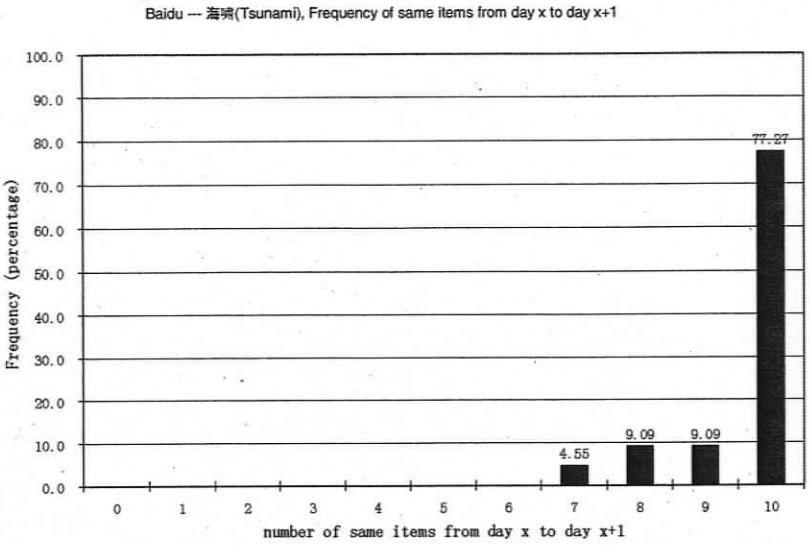


Figure 5.7 *Histogram for daily duplication frequency: Baidu-Tsunami*

Group DD-U: There is no dominant bar at either ends, rather, the bars are distributed almost uniformly along the X-axis with similar heights. In this case, there are various degrees of duplication between consecutive days. An example of this is shown in Figure 5.5 for the keyword Chao Nv using search engine Zhongsou.

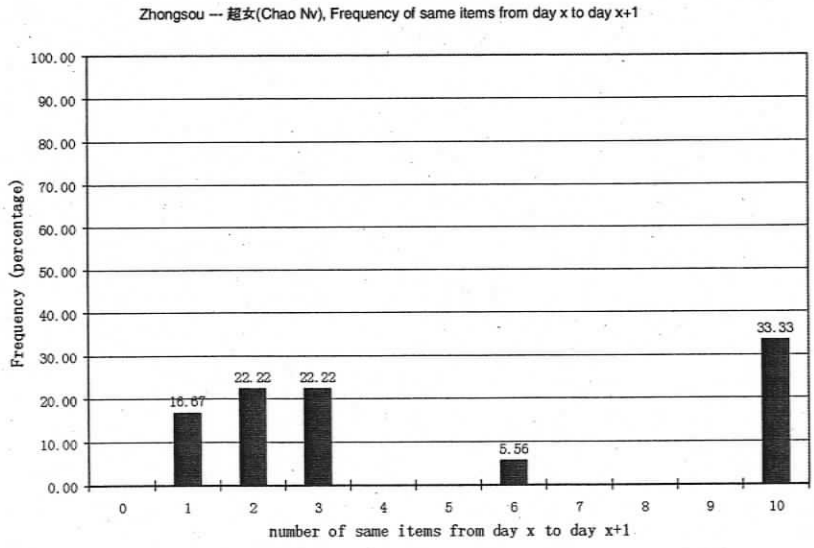


Figure 5.8 *Histogram for daily duplication frequency: Zhongsou: Chao Nv*

Group DD-L: The highest bars are concentrated at the left side. An example of this is shown in Figure 5.6 for search engine Tianwang with keyword Hurricane. In this case duplication between consecutive days is relatively low.

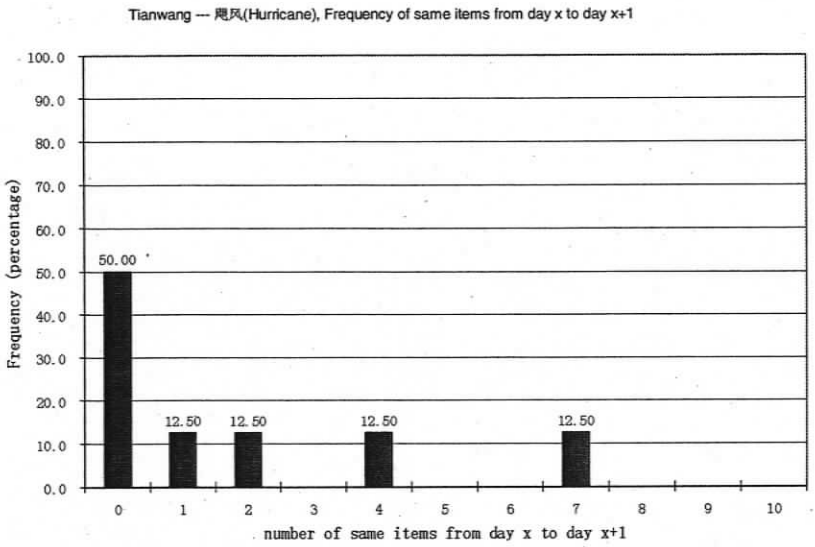


Figure 5.9 *Histogram for daily duplication frequency: Tianwang-Hurricane*

Possible reason for the different shapes:

We speculate that the main reason for the different shapes in the histograms is due to the variations in the update frequency of a search engine's database. If a search engine updates its database on a regular basis, a histogram belongs to DD-L is expected. On the contrary, if the database update frequency is low (weekly for example) the histogram would belong to DD-R. Histogram belonging to DD-U represents a situation in between these two cases.

The query keyword also has an effect on the search results. When a keyword is related to a widely discussed and recent topic, there are lots of changes almost on an hourly basis. Therefore, the search results are very different between consecutive days and a DD-L histogram is expected. For a topic which is relatively outdated, the number of updates would be less frequent. As a result, the histogram may look like a DD-U, or in between DD-U and DD-R, or in between DD-U and DD-L. An example is the daily duplication frequency histograms for Google in 2005 and 2006 as shown in Table 5.3. Tsunami was a very hot topic in 2005 and therefore the histogram for this time period is a

DD-R. In 2006, Tsunami was still a hot keyword but less popular than in 2005. Therefore, the 2006 histogram looks different. It can still be considered as a DD-R but is becoming closer to DD-U indicating less Tsunami related updates on the web.

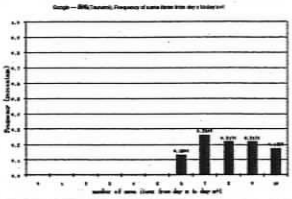
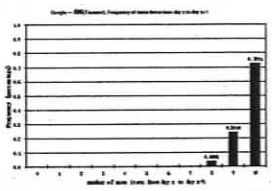
2005-9	2006-9
	
Group DD-R	Group DD-R

Table 5.3 *Comparison for Google-Tsunami between 2005 and 2006*

Another reason for the change of shape may be related to the ranking algorithm used by the search engine. It has been shown in previous work [22] that sometimes a search engine fails to rank web pages in a consistent manner. The ranking of some pages keep changing over time. These "mishandled URLs" can also affect the shape of the histograms.

5.4.2 Period Duplication Frequency

The histograms for period duplication frequency can also be classified into three groups.

Group PD-R: The bars at the right end have comparatively higher height. Figure 5.7 shows two such cases. In Figure 5.7 (a) the bar at the right end is high and the bars to its left are much lower. This means many items are repeated from one day to another. An extreme case is shown in Figure 5.7 (b) where the ten items are repeated for everyday during the data collection period.

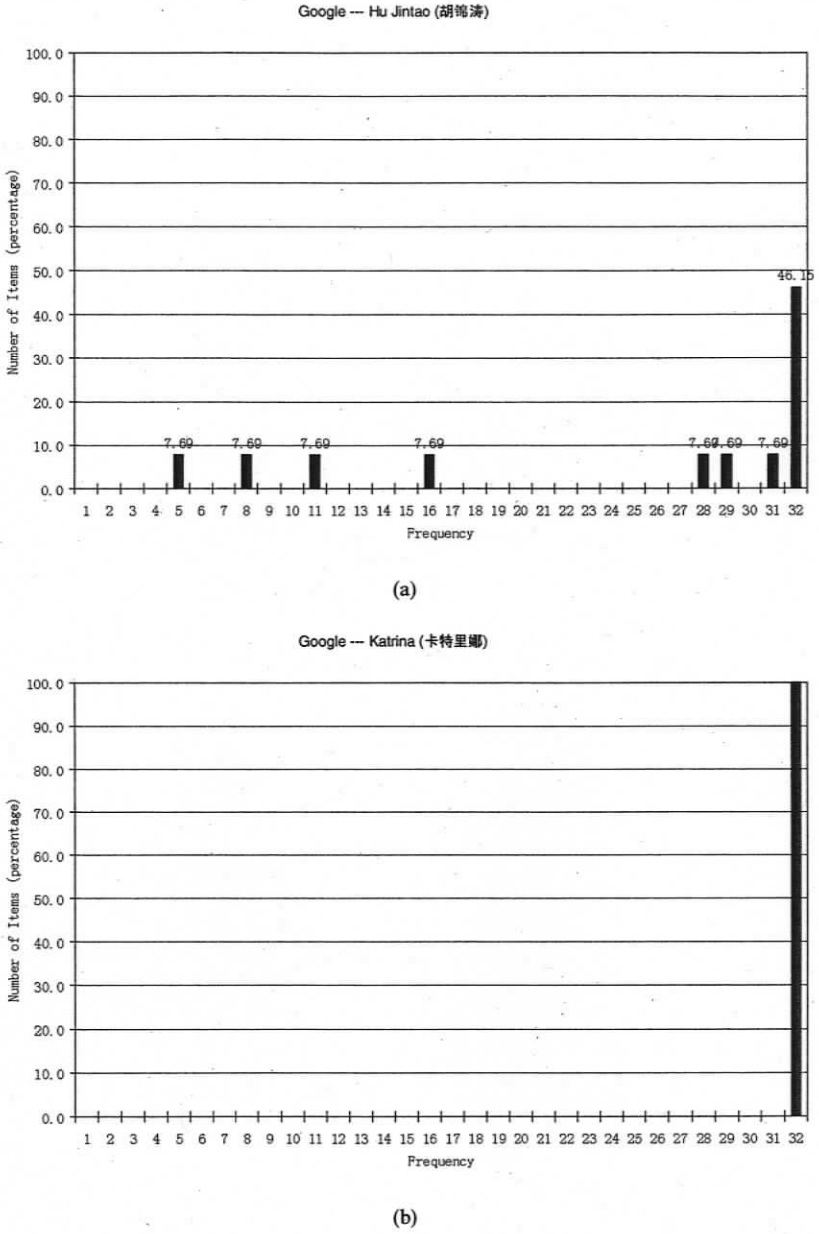


Figure 5.7 *Histogram for period duplication frequency: Google-(a) Hu Jintao, (b)Katrina*

Group PD-L: The highest bars are concentrated at the left side. Figure 5.8 shows an example of Google with keyword Rita (the hurricane Rita that hit the U.S. Gulf Coast in 2005).

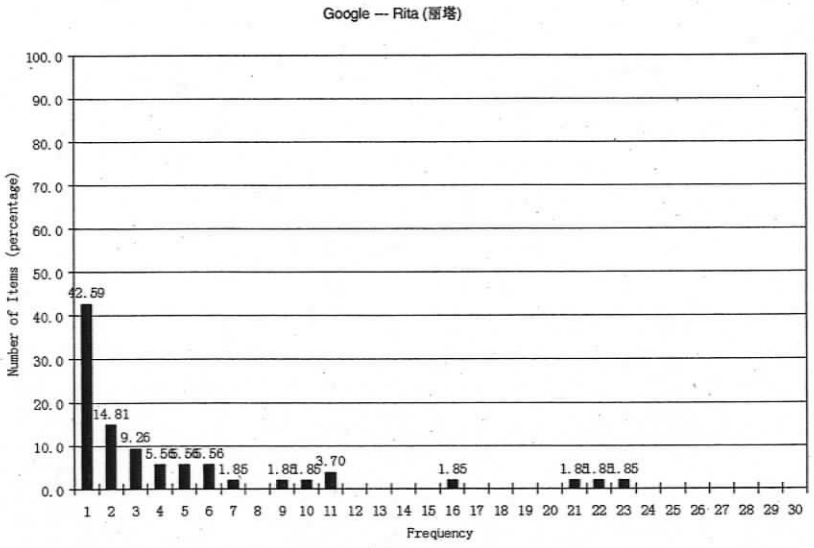


Figure 5.8 *Histogram for period duplication frequency: Google-Rita*

Group PD-M: The bars in the middle are the highest. This represents a situation between PD-R and PD-L. There are many items which appear during half of the data collection period but are not on the list all the time. An example is shown in Figure 5.9.

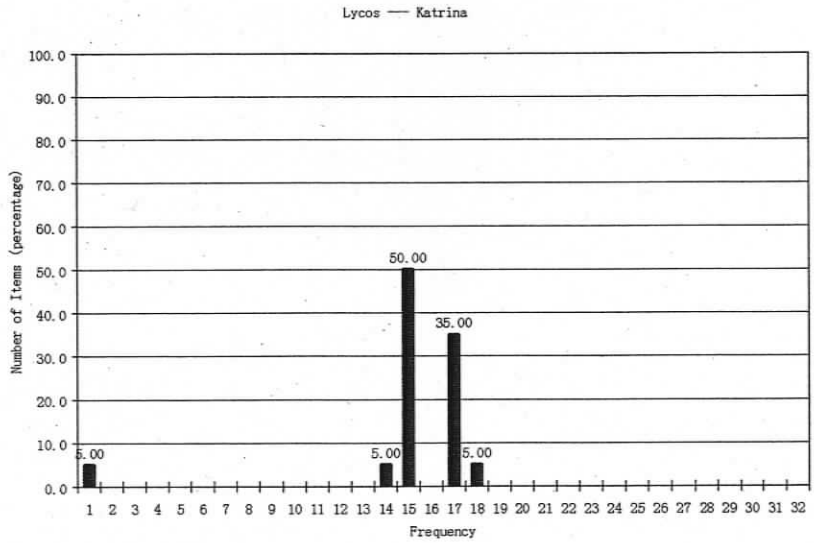


Figure 5.9 *Histogram for period duplication frequency: Lycos-Katrina*

Possible reason for the different shapes:

The shapes to a great extent depend on the search engine's database update frequency and the popularity of the keyword. If the database is updated frequently and the keyword is a

very hot topic, there will be many new items and very few repetitions, resulting in a PD-R histogram. If the topic is outdated and/or the database is updated infrequently, a PD-L histogram is most probable. Group PD-M represents situation in between these two extremes.

5.4.3 Daily Rank Change Frequency

One of the most important properties of any search engine is its ability to rank the relevant pages properly. The shape of the rank-change histogram gives very important information regarding the search engine. Again, three groups of histograms are identified.

Group DR-N: Bar # is very high as shown in Figure 5.10. This means that there are many new items on the top ten list from day x to day x+1.

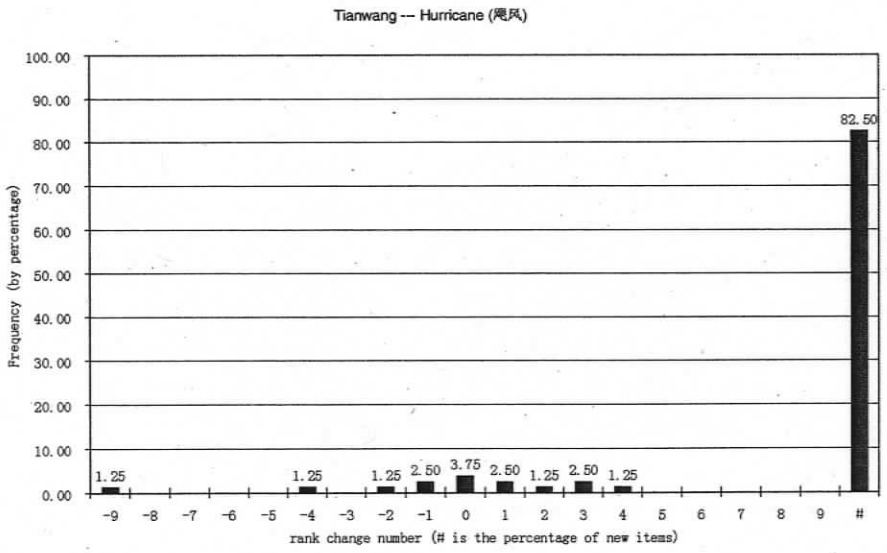


Figure 5.10 *Example of very high Bar #: Tianwang-Hurricane*

Group DR-Z: Bar 0 is very high as shown in Figure 5.11. This means that there is very little rank change from day x to day x+1.

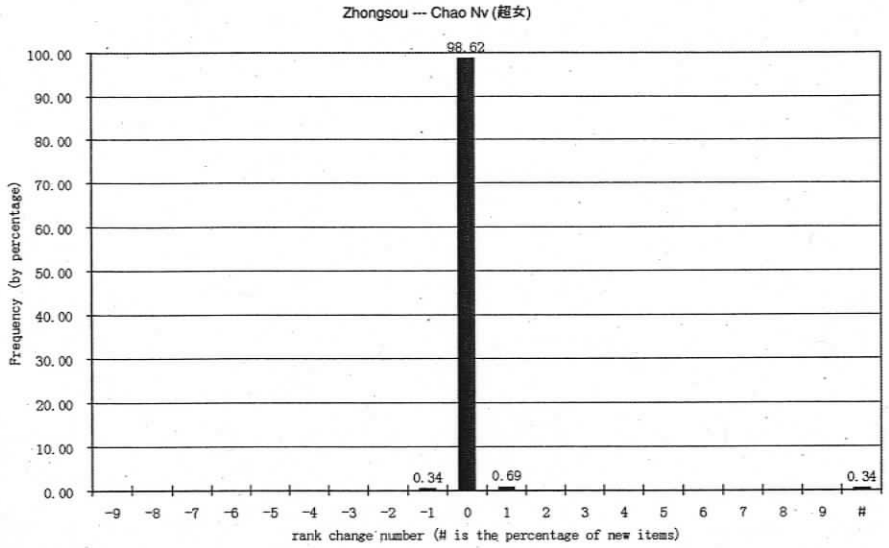


Figure 5.11 *Example of very high Bar 0: Zhongsou-Chao Nv*

Group DR-N-Z: Both bar # and bar 0 have relatively high height as shown in Figure 5.12. This means that there are many new items in consecutive days, but the common items have no change in their ranking.

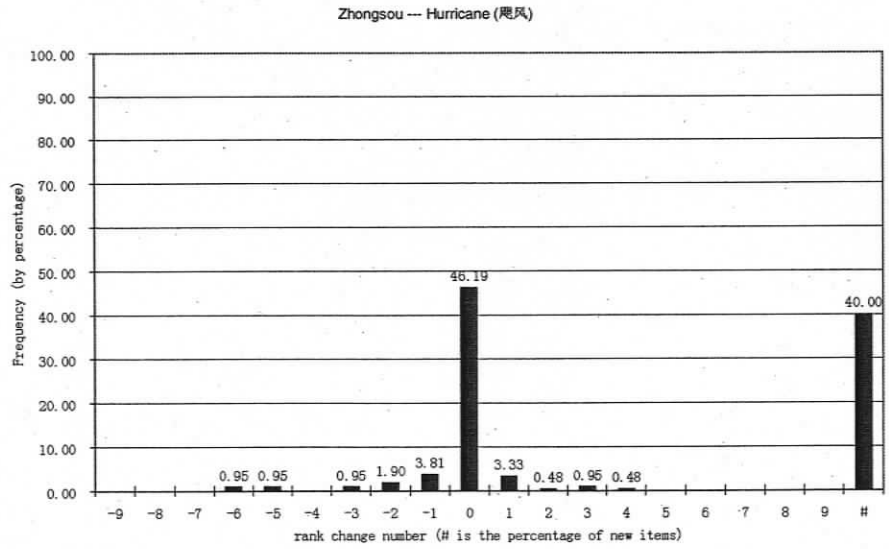


Figure 5.12 *Example of Bar # and Bar 0 both have high height: Zhongsou-Hurricane*

Possible reasons for the different shapes.

It is almost impossible to conclude the specific reasons for the different shapes because there are so many factors that could affect the ranking: the ranking algorithm used, how often the search engine changes its ranking algorithm, and the database update frequency.

5.4.4 Concluding Remarks

From the above analyses, one can conclude that there are many factors contribute to the observed characteristics of the performance measurement metrics:

- The ranking algorithm of the search engine is modified.
- The database update frequency has changed.
- The search engine has found pages which are more relevant than the existing pages in its database.
- The search engine discovers that the content of the page has changed and it is no longer relevant.
- The URL disappears from the web or the crawler fails to find the same page because of communication problem or server failure.
- Some unknown reasons.

5.5 A Quantitative Method for Comparing and Classifying Histograms

The previous sections have presented the performance measurement metrics with the aid of histograms. The histograms are classified into different groups by visual inspection; however, this manual process is not practical. This section introduces a method to compare and classify histograms in a quantitative and deterministic manner. The proposed method is similar to that used for image processing [57] but is novel in its application to search engine evaluation.

Histograms are statistical distributions and therefore can be characterized by a number of parameters such as mean, median, skewness, etc. A vector can be used to represent a histogram. Let this vector be $h_i=\{p_{i1},p_{i2}, p_{i3}, ...\}$ where p_i's are the various statistical parameters for a specific histogram.

Since histograms can be represented by vectors, therefore different histograms can be compared by using their corresponding vectors. The most common way to compare two vectors is to compute the "distance" between them. However, there is one significant difference between an ordinary vector v_i (representing the three-dimensional velocity of an object for example) and the vector h_i representing a histogram. In the former case each component of the vector represents similar quantity and it is logical to

assign similar weight to the different components. Therefore, a measure of distance between two vectors $v1=\{x1, y1, z1, ...\}$ and $v2=\{x2, y2, z2, ...\}$ is given by

$$\Delta = (x1-x2)^2 + (y1-y2)^2 + (z1-z2)^2 + ... \quad (5.3)$$

In the case of h_i, the magnitude of the components p_{i1}, p_{i2}, p_{i3} etc. can be significantly different since they represent different properties of the histograms. For example, the median is usually larger than the standard deviation. Therefore it is reasonable to scale the different components by appropriate weights. This will make the comparison more meaningful as all the factors are equally relevant. The same approach has been taken by previous work in other fields as well [57].

Suppose h_1 and h_2 are two vectors representing two histograms.

$$h_1 = \{p_{11}, p_{12}, p_{13}, ...\} \quad (5.4)$$

$$h_2 = \{p_{21}, p_{22}, p_{23} ..... \} \quad (5.5)$$

Based on the description above, the distance between these two vectors (Δ_{1-2}) can be calculated as

$$\Delta_{1-2} = w_1(p_{11}-p_{21})^2 + w_2(p_{11}-p_{21})^2 + w_3(p_{11}-p_{21})^2 + ... \quad (5.6)$$

Here, w_i's are the weights assigned to different parameters to make their significance similar.

Three commonly used statistical properties, median (μ), standard deviation (σ) and skewness (S) are used in our method. Mean is not used since the histograms are normalized. So the sum of the heights of the bars in any histogram always equals to 1, rendering this measure useless. For a data set $\{x_1, x_2, ..., x_N\}$ the standard deviation σ and skewness S_k are defined as

$$\sigma = [\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu)^2]^{1/2} \quad (5.7)$$

$$S_k = \frac{1}{N}\sum_{i=1}^{N}(\frac{x_i-\mu}{\sigma})^3 \quad (5.8)$$

Though median, standard deviation and skewness give important information about the histogram, to complete the description some additional statistical parameter is necessary. The concept of expected value is used here to include the bars' values on the X- and the Y-axis for consideration. For histograms within the same group, their

expected values should be similar. The ExpectedValue is calculated according to the following equation:

$$ExpectedValue = \sum_{i=1}^{n} w_i * Bar_i \tag{5.9}$$

Here, n is the number of bars, w_i is the weight for Bar_i, and Bar_i is the value for the i^{th} bar. The values of the weights (w_i) for the histograms are $w_i=k$, with $k = 1$ to 11 for the daily duplication frequency, $k = 1$ to 20 for the daily rank change frequency, and $k = 1$ to T for the period duplication frequency.

The proposed method starts with identifying the benchmark histograms for different groups and determining the values of the statistical parameters for them. Tables 5.4, 5.5 and 5.6 show the benchmark histograms for the different groups in daily duplication frequency, period duplication frequency and rank change frequency, respectively. These histograms resemble the "ideal" or extreme cases for the different groups. The relevant statistical parameters for the benchmark histograms are shown in Tables 5.7, 5.8 and 5.9.

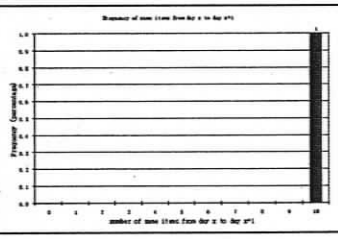
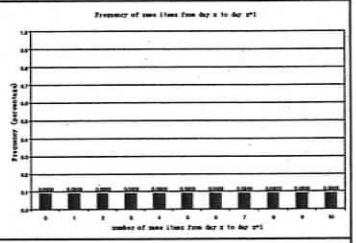
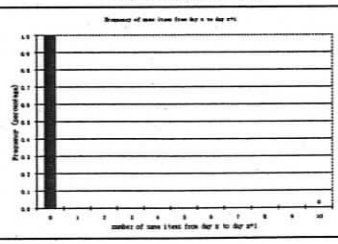
Daily Duplication	Group DD-R	
	Group DD-U	
	Group DD-L	

Table 5.4 *Benchmark histograms for daily duplication frequency*

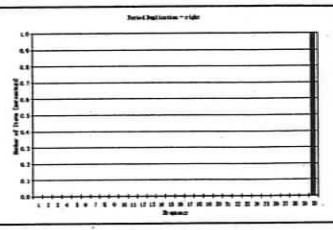
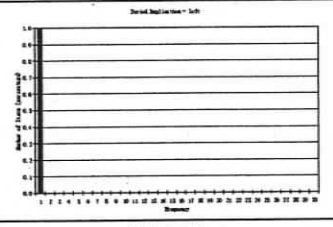
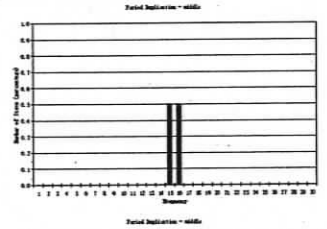
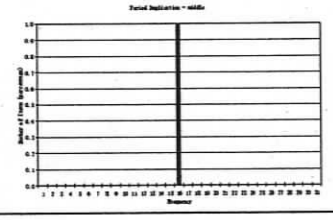
Period Duplication	Group PD-R	
	Group PD-L	
	Group PD-M	 

Table 5.5 *Benchmark histograms for period duplication frequency*

For Group PD-M, if the frequency scale x_1 is an even number, the two-bar benchmark is used. If the frequency scale x_2 is an odd number (here, $x_2=x_1+1$), the one-bar benchmark is used. The values of the four statistical parameter values are the same for these two benchmarks.

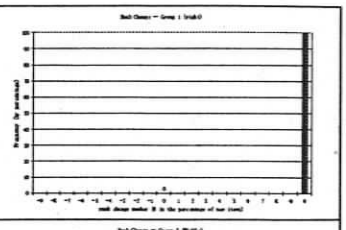
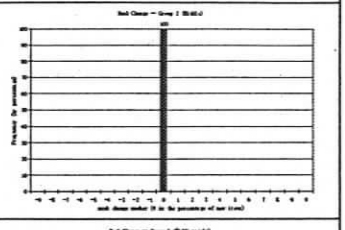
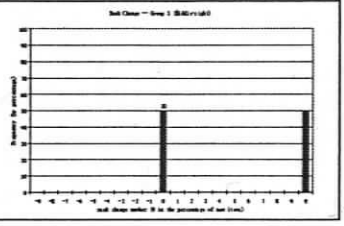
Rank Change	Group DR-N	
	Group DR-Z	
	Group DR-N-Z	

Table 5.6 *Benchmark histograms for daily rank change frequency*

	Median	Standard Deviation	Skewness	ExpectedValue
Group DD-R	0	0.30	3.32	11
Group DD-U	0.09	0	-0.66	6
Group DD-L	0	0.30	3.32	1

Table 5.7 *Statistical parameter values for the daily duplication frequency benchmark*

	Median	Standard Deviation	Skewness	ExpectedValue
Group PD-R	0	0.18	5.48	31
Group PD-L	0	0.18	5.478	1
Group PD-M	0	0.13	3.66	15.5

Table 5.8 *Statistical parameter values for the period duplication frequency benchmark*

	Median	Standard Deviation	Skewness	ExpectedValue
Group DR-N	0	0.22	4.47	20
Group DR-Z	0	0.22	4.47	10
Group DR-N-Z	0	0.15	2.89	15

Table 5.9 *Statistical parameter values for the daily rank change frequency benchmark*

The next step is to determine the weights necessary to scale the statistical parameters to similar magnitude. Through trial and error, the most appropriate weights used for the histograms are determined and shown in Table 5.10.

	Median	Standard Deviation	Skewness	ExpectedValue
Daily duplication frequency	1000	100	1	1
Period duplication frequency	1	100	1	0.1
Daily rank change frequency	1	0.01	1	1

Table 5.10 *The weights used for different kinds of histograms*

To determine the distance between a histogram and a benchmark, the following equation is used:

$$\Delta_{engine-standard} = w_1(\mu_{eng} - \mu_{bench})^2 + w_2(\sigma_{eng} - \sigma_{bench})^2 + w_3(S_{eng} - S_{bench})^2 + w_4(\chi_{eng} - \chi_{bench})^2 \quad (5.10)$$

Here μ_{eng}, σ_{eng}, S_{eng}, χ_{eng} are the median, standard deviation, skewness and ExpectedValue of the histogram under investigation and μ_{bench}, σ_{bench}, S_{bench}, χ_{std} are the median, standard deviation, skewness and ExpectedValue of the benchmark. The classification of a histogram is based on the shortest distance from a benchmark.

5.6 Validation of the Proposed Method

To validate our approach, a number of cases are presented here. It should be mentioned that the example cases were chosen at random from the collected data and they represent typical results obtained for all keywords and search engines. For the first two cases, it is very easy to determine by visual inspection which group the histogram belongs to. The proposed method was applied to many histograms obtained from the collected data sets. The results of the distance measure using statistical parameters always match that obtained from visual inspection. The third example is for a case where it is difficult to classify the histogram with confidence solely by visual inspection and therefore illustrates the usefulness of the proposed method.

The first example is for English search engines and the keyword is Tsunami. Data collection period is from 2006-9-12 to 2006-11-13. Tables 5.11, 5.12 and 5.13 show the results for the three performance measurement metrics.

Keyword / Date / Metric	Parameters	Google	Yahoo	MSN	Lycos	Hotbot	AOL
Tsunami / 2006-09-12 to 2006-11-13 /Daily duplication frequency	Median	0.00	0.00	0.00	0.00	0.00	0.00
	Standard Deviation	27.00	29.10	18.07	24.05	19.93	28.08
	Skewness	3.25	3.31	2.42	3.17	2.62	3.29
	Expected Value	10.90	9.97	10.35	10.55	10.35	10.94
	Distance from Group DD-R	7.14	9.36	4.37	5.86	4.76	7.72
	Distance from Group DD-U	54.83	48.24	39.10	49.42	41.94	56.11
	Distance from Group DD-L	105.14	88.71	91.47	96.83	91.85	106.43
	Group to which the engine belongs	Group DD-R	Group DD-R	Group DD-R	Group DD-R	Group DD-R	Group DD-R

Table 5.11 *Daily duplication frequency: Tsunami*

Keyword/ Date/ Metric	Parameters	Google	Yahoo	MSN	Lycos	Hotbot	AOL
Tsunami / 2006-09-12 to 2006-11-13 / Period duplication frequency	Median	0.00	0.00	0.00	0.00	0.00	0.00
	Standard Deviation	0.14	0.14	0.07	0.10	0.08	0.15
	Skewness	5.28	5.41	3.66	3.92	3.03	5.46
	Expected Value	25.83	28.80	19.50	16.74	15.14	29.09
	Distance from Group PD-R	2.98	0.66	18.25	24.02	33.09	0.50
	Distance from Group PD-L	61.98	77.46	39.25	28.45	27.95	79.05
	Distance from Group PD-M	12.91	20.33	1.85	0.25	0.83	21.29
	Group to which the engine belongs	Group PD-R	Group PD-R	Group PD-M	Group PD-M	Group PD-M	Group PD-R

Table 5.12 *Period duplication frequency: Tsunami*

Keyword / Date/ Metric	Parameters	Google	Yahoo	MSN	Lycos	Hotbot	AOL
Tsunami / 2006-09-12 to 2006-11-13 / Daily rank change frequency	Median	0.00	0.00	0.00	0.00	0.00	0.00
	Standard Deviation	19.44	22.28	17.71	20.63	20.12	20.64
	Skewness	4.43	4.47	4.33	4.45	4.44	4.45
	Expected Value	10.10	10.04	10.37	10.41	10.60	10.07
	Distance from Group DR-N	101.81	104.15	97.84	96.24	92.37	102.85
	Distance from Group DR-Z	3.88	4.86	5.25	4.36	4.45	4.21
	Distance from Group DR-N-Z	30.07	32.05	26.57	27.75	25.71	30.98
	Group to which the engine belongs	Group DR-Z	Group DR-Z	Group DR-Z	Group DR-Z	Group DR-Z	Group DR-Z

Table 5.13 *Daily rank change frequency: Tsunami*

Table 5.14 shows the corresponding histograms for the results shown in Tables 5.11, 5.12, and 5.13.

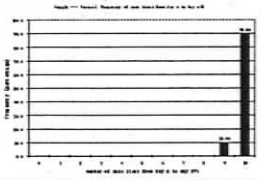
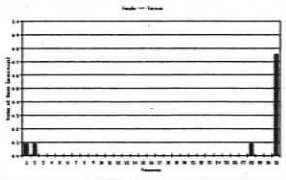
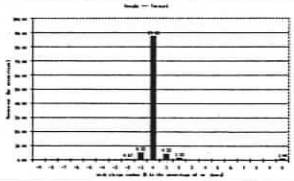
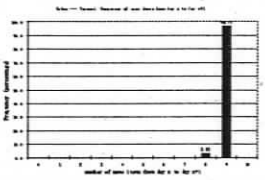
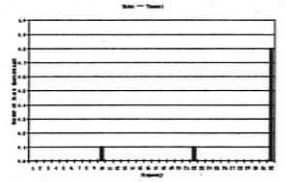
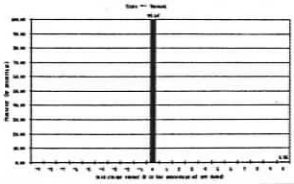
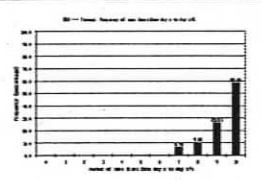
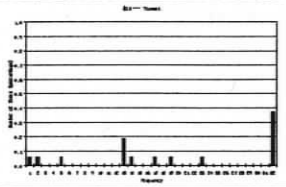
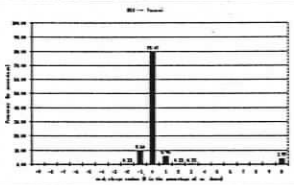
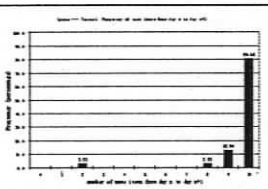
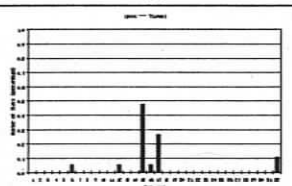
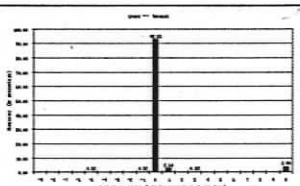
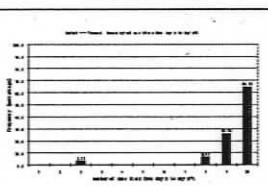
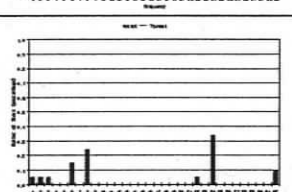
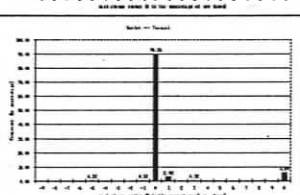
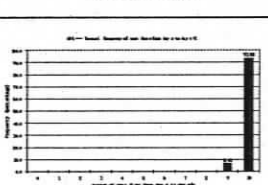
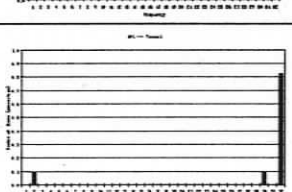
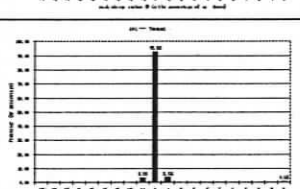
Tsunami	Daily duplication frequency	Period duplication frequency	Daily rank change frequency
Google			
Yahoo			
MSN			
Lycos			
Hotbot			
AOL			

Table 5.14 *Corresponding histograms for table 5.11, 5.12 and 5.13*

The second example is for Chinese search engines and the search word is New Orleans. Data collection period is from 2005-09-20 to 2005-10-25. Tables 5.15, 5.16, and 5.17 summarize the results for the three performance measurement metrics.

Keyword/ Date/ Metric	Parameters	Google	Yahoo	Baidu	Tian wang	Zhong sou
New Orleans / 2005-09-20 to 2005-10-25 / Daily duplication frequency	Median	0.00	0.00	0.06	0.11	0.08
	Standard Deviation	0.12	0.15	0.09	0.10	0.11
	Skewness	1.17	1.95	1.18	0.41	1.93
	Expected Value	8.94	9.12	7.53	5.89	7.62
	Distance from Group DD-R	12.10	7.58	24.61	51.10	22.86
	Distance from Group DD-U	21.73	27.19	7.54	2.51	10.77
	Distance from Group DD-L	70.92	69.93	55.20	48.88	55.16
	Group to which the engine belongs	Group DD-R	Group DD-R	Group DD-U	Group DD-U	Group DD-U

Table 5.15 *Daily duplication frequency: New Orleans*

Keyword/ Date/ Metric	Parameters	Google	Yahoo	Baidu	Tian wang	Zhong sou
New Orleans / 2005-09-20 to 2005-10-25 / Period duplication frequency	Median	0.04	0.02	0.01	0.06	0.02
	Standard Deviation	0.07	0.07	0.10	0.19	0.12
	Skewness	1.56	1.04	2.60	2.66	2.59
	Expected Value	7.12	7.22	3.27	2.33	2.98
	Distance from Group PD-R	73.59	77.57	85.79	90.15	87.20
	Distance from Group PD-L	20.31	24.87	9.43	8.10	9.07
	Distance from Group PD-M	11.73	14.08	16.13	18.72	16.82
	Group to which the engine belongs	Group PD-M	Group PD-M	Group PD-L	Group PD-L	Group PD-L

Table 5.16 *Period duplication frequency: New Orleans*

Keyword/ Date/ Metric	Parameters	Google	Yahoo	Baidu	Tian wang	Zhong sou
New Orleans / 2005-09-20 to 2005-10-25 / Daily rank change frequency	Median	0.00	0.00	0.00	0.00	0.00
	Standard Deviation	13.82	14.49	7.23	6.56	12.80
	Skewness	4.12	4.05	2.79	4.08	4.34
	Expected Value	12.31	11.25	14.14	15.46	13.22
	Distance from Group DR-N	61.11	78.69	37.63	21.21	47.64
	Distance from Group DR-Z	7.30	3.79	20.45	30.32	11.95
	Distance from Group DR-N-Z	10.63	17.44	1.25	2.05	6.80
	Group to which the engine belongs	Group DR-Z	Group DR-Z	Group DR-N-Z	Group DR-N-Z	Group DR-N-Z

Table 5.17 *Daily rank change frequency: New Orleans*

Table 5.18 shows the histograms from different search engines for New Orleans. A comparison of these histograms with the results presented in Tables 5.15, 5.16, and 5.17 shows exact classifications for daily duplication frequency and period duplication frequency. For daily rank change frequency, the classifications from the proposed method match that with visual inspection, except for Google and Zhongsou. When the distance measures show significant difference to different benchmarks, then the classification can be done easily by inspection. However, if the distances to the benchmarks are similar and comparable, let say 5 or less, then classification by visual inspection becomes very subjective and difficult.

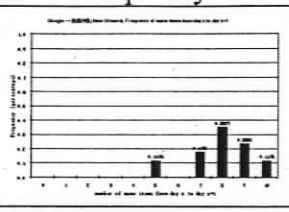
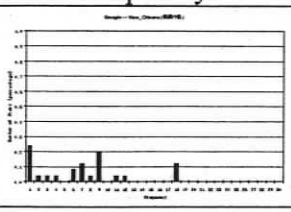
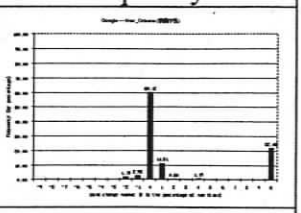
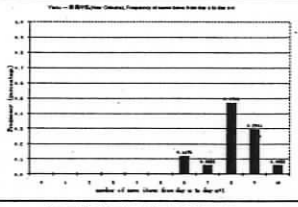
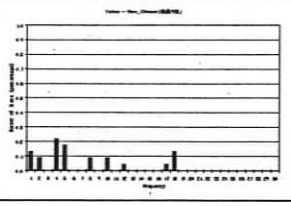
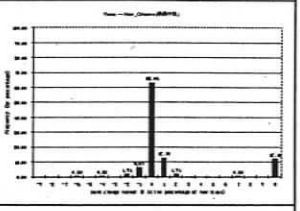
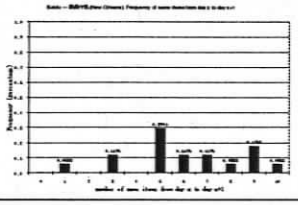
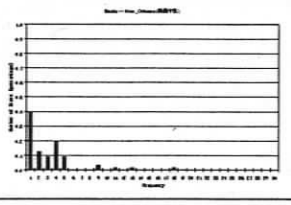
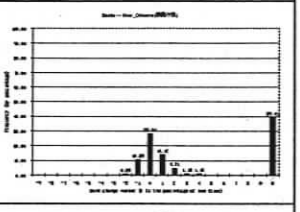
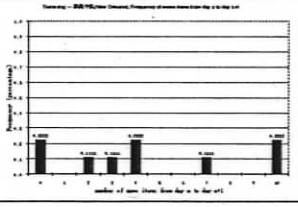
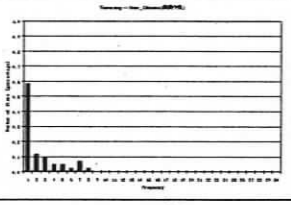
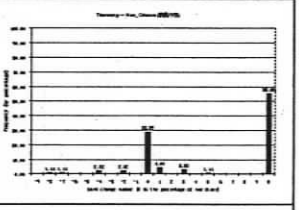
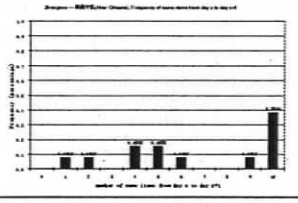
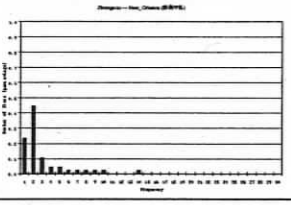
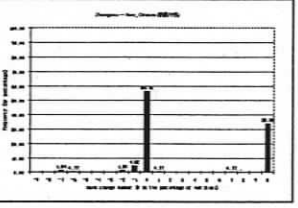
New Orleans	Daily duplication frequency	Period duplication frequency	Daily rank change frequency
Google			
Yisou			
Baidu			
Tianwang			
Zhongsou			

Table 5.18 *Corresponding histograms for tables 5.15, 5.16 and 5.17*

To illustrate the applicability of the proposed method, the classification of the daily rank change histogram for Google is examined more closely here. It is difficult to determine whether the histogram belongs to group DR-N-Z or group DR-Z by visual inspection, even with an enlarged histogram as shown in Figure 5.13. However, our proposed method provides a deterministic means that classifies the histogram under Group DR-Z (distance 7.3) rather than DR-N-Z (distance 10.6), thus eliminating any ambiguity, subjectivity, and uncertainty associated with human inspection.

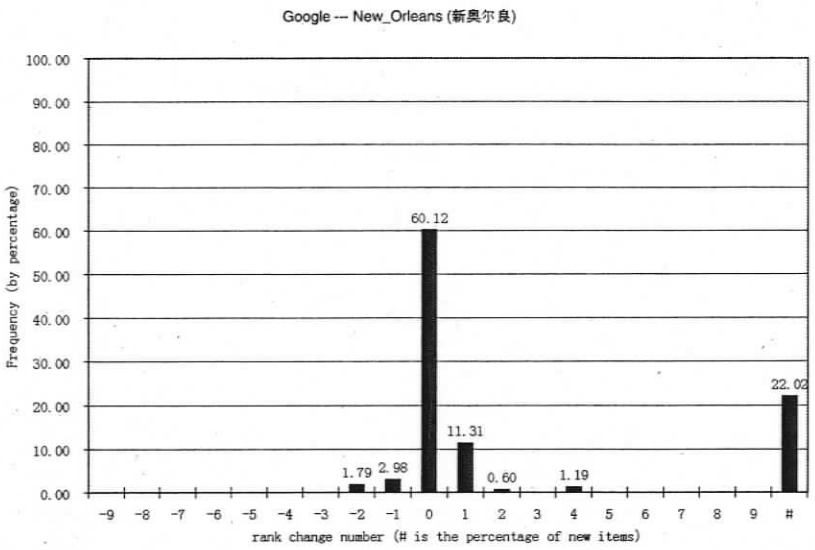


Figure 5.13 *Histogram for daily rank change frequency: Google-New Orleans*

5.7 Application of the Proposed Classification

The proposed classification method identifies the benchmarks for the different groups and compares histograms for specific keyword and search engine to those benchmarks. A user can quickly find out which group a histogram belongs to and get an idea of the characteristics of that search engine. Furthermore, knowing the characteristics of the search results for a particular topic enables a user to choose the most appropriate search engine that suits his/her needs.

5.8 Conclusions

The variation of search results over time for a number of English and Chinese search engines are investigated using the three performance measurement metrics. The histograms representing the results are classified into different groups according to their shapes, patterns, and heights. Possible reasons for these groupings are discussed. A deterministic method to compare and classify histograms quantitatively is presented. The proposed method, which can easily be automated, is a very useful tool for analyzing and comparing results from different search engines.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

The main objective of this thesis is to explore evaluation methods for search engines and to develop a model for automated evaluation. The main contributions of this thesis are:

- A novel model for search engine evaluation is proposed. Search engine features are divided into groups and subgroups. This provides a very convenient way to compare search engines, in whole or in particular aspects. The concept of a common list is introduced to make automation of the evaluation model possible. The model is used in the evaluation of several search engines to illustrate its usefulness for researchers, individual users, and service providers.
- An extensive survey of the existing Chinese search engines has been carried out. The unique challenges that Chinese search engines are facing are examined. The most popular engines are identified and their performance is analyzed using different evaluation methods.
- The variation of search results over time is investigated. Three performance measurement metrics are introduced to quantify the variation. Data are collected for several popular English and Chinese search engines. The metrics derived from the collected results are represented by histograms and classified into different groups based on the shape of the histograms. Possible reasons for the different shapes are discussed.
- A method is proposed for quantitative comparison of the histograms based on their statistical characteristics. This provides a fully automated means to compare histograms and removes the uncertainty and inaccuracy that may result from visual inspection.

6.2 Future Works

Further investigation can be undertaken to improve the proposed evaluation model, performance measurement metrics, and classification of search results. Possible addition and modification of the feature parameters would make the evaluation model more complete and ensure more meaningful evaluation. The evaluation results shown in this work are for the purpose of proof of concept only. More research is necessary to make the model and the evaluation methods practical.

The usefulness of the proposed evaluation model depends critically on the weights assigned to the different features and on the quality of the common list. Finding a proper scheme to assign appropriate weights to different features is important, though this weight assignment could be a personal preference given by individual users. The common list used in this work is generated by simply merging the search results from different engines. Further investigation is necessary to ascertain whether a better fusion algorithm will provide more meaningful and useful information on the search engines being evaluated.

Bibliography

[1] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomlins, J. Wiener, "Graph structure in the Web", in proceedings of the 9th international World Wide Web conference, May 2000, Computer Networks and ISDN Systems, 33, pp. 309–320.

[2] A. Chowdhury, I. Soboroff, "Automatic evaluation of world wide web services", Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, Tampere, Finland, August 11 - 15, 2002.

[3] A. Gulli, A. Signorini, « Building an open source metasearch engine", in 14th WWW, 2005.

[4] B. J. Jansen, A. Spink, J. Bateman, T. Saracevic, "Real life information retrieval: A study of user queries on the Web", SIGIR Forum, 32(1), pp. 5-17, 1998.

[5] C. Dwork, R. Ku,ar, M. Naor, D. Sivakumar, "Rank aggregation methods for the web", Proceedings of the 10th World Wide Web Conference, Hong Kong, pp. 613-622, May 2001.

[6] China Internet Information Center, "A Survey and Report on the Status of Internet Development in China", available at (Mar. 9, 2008): http://www.cnnic.net.cn/download/2004/2004072002.pdf

[7] N.-P. Chen, et al, "Chinese Mac Character Sets and Encodings", available at (Sep. 21, 2004): http://www.yale.edu/chinesemac/pages/charset_encoding.html

[8] C. Oppenheim, A. Morris, C. McKnight, "The evaluation of WWW search engines", Journal of Documentation, 56 (1), pp. 71-90, 2000.

[9] C. Silverstein, M. Henzinger, J. Marais, M. Moricz, „Analysis of a very large Alta Vista query log", Technical Report 1998-014, COMPAQ Systems Research Center, Palo Alto, Ca, USA, 1998.

[10] Chinese-search-engine.com, "Marketing China: Simple Facts About China", available at (Oct. 22, 2006): http://www.internetworldstats.com/asia.htm

[11] D. Hawking, P. Bailey, K. Griffiths, "Measuring search engine quality", Information Retrieval, 4, pp. 33-59, 2001.

[12] D. Hawking, "Web Search Engines: Part 1", IEEE Computer, pp. 86-88, Jun. 2006.

[13] D. Lewandowski, H. Wahlig, G. Meyer-Bautor , "The Freshness of Web search engines' databases", Journal of Information Science, 2006.

[14] D. Sullivan, Search Engine Watch Newsletter, Jan 28, 2005, available at (Mar. 9, 2008): http://searchenginewatch.com/showPage.html?page=2156481

[15] E. M. VOORHEES, D. K. HARMAN, *Text Retrieval Conferences,* 2002, available at (Mar. 9, 2008): http://trec.nist.gov/.

[16] S. Foo, H. Li, "Chinese Word Segmentation and Its Effect on Information Retrieval", Information Processing and Management, vol. 40, issue 1, pp. 161-190, Jan. 2004.

[17] F.W. Lancaster, E.G. Fayen. "Information retrieval: on-line", Los Angeles, CA: Melville Pub., Information Sciences Series 1973.

[18] H. Chu, M. Rosenthal, "Search Engines for the World Wide Web: A Comparative Study and Evaluation Methodology", ASIS 1996 Annual Conference, October 19-24, 1996.

[19] H. Leighton, J. Srivastava, "Precision among WWW search services (search engines): AltaVista, Excite, HotBot, Infoseek and Lycos" 1997, available at (Jun. 11, 2005): http://www.winona.edu/library/webind2.htm.

[20] Internet World Stats Usage and Population Statistics, "Asia Marketing Research, Internet Usage, Population Statistics and Information", available at (Sep. 21, 2004): http://chinese-search-engine.com/chinese-search-engine/survey.htm

[21] J. Allan, et al, "Challenges in Information Retrieval and Language Modeling", Report of a Workshop held at the Center for Intelligent Information Retrieval, University of Massachusetts Amherst, Sep. 2002.

[22] J. Bar-Illan, "Methods for measuring search engine performance over time", Journal of American Society for Information Science and Technology, 53(4), pp. 308-319, 2002.

[23] J. Bar-Illan, "Criteria for Evaluating Information Retrieval Systems in Highly Dynamic Environments", 2nd International Workshop on Web Dynamics in conjunction with the 11th International World Wide Web Conference Honolulu, Hawaii, USA, May 7, 2002, pp. 70-77.

[24] J. Bar-Illan, M. Levene, M Mat-Hassan, "Dynamics of search engine rankings- A case study", Proceedings of the 3rd International Workshop on Web Dynamics. May 2004.

[25] J. Bar-Illan, M Mat-Hassan, M. Levene, "Methods for comparing rankings of search engine results", Computer Networks 50, pp. 1448–1463, 2006.

[26] A. Kingoff, "Comparing Internet Search Engines", IEEE Computer, pp. 117-118, Apr. 1997.

[27] M. Ljosland, "A Comparison between Twenty Web Search Engines on Ten Rare Words", available at (Mar. 9, 2008): www.aitel.hist.no/~mildrid/dring/paper/Comp20.doc

[28] R.W.P. Luk, K.L. Kwok, "A Comparison of Chinese Document Indexing Strategies and Retrieval Models", ACM Transactions on Asian Language Information Processing, vol. 1, no. 3, pp. 225-268, Sep. 2002.

[29] L. T. Su, H. L. Chen, X. Y. Dong, "Evaluation of Web-based search engines from an end-user's perspective: A pilot study", Proceedings of the 61st Annual Meeting of the American Society for Information Science, Pittsburgh, PA., pp. 348-361, 1998.

[30] W.-C. Lin, H.-H. Chen, "Description of NTU Approach to NTCIR3 Multilingual Information Retrieval", Proceedings of the Third NTCIR Workshop, Sep. 2001.

[31] M. E. Keen, "Evaluation parameters", in G. Salton (Ed.), The SMART retrieval system—Experiments in automatic document processing. Englewood Cliffs, NJ: Prentice-Hall., 1971.

[32] M. Gordon and P. Pathak, "Finding information on the world wide web: the retrieval effectiveness of search engines", IP&M, 25(2), pp. 141-180, 1999.

[33] M. H. Chignell, J. Gwizdka, R. C. Bodner, "Discriminating meta-search: A framework for evaluation", Information processing and management, 35(3), pp. 337-362, 1999.

[34] M. Kobayashi, K.Takeda "Information retrieval on the Web", ACM Computing Surveys, 32(2), pp. 144-173, 2000.

[35] L. Ma, "An Introduction and Comparison of Chinese Search Site", eSAS World, 139-146, Jul. 1998, (in Chinese), available at (Mar. 9, 2008): http://www.mypcera.com/softxue/txt/s35.htm

[36] M. R. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork, "Measuring Index Quality Using Random Walks on the Web", available at (Jun 21, 2007): http://www8.org/w8-papers/2c-search-discover/measuring/measuring.html

[37] M. W. Berry, M. Browne, "Understanding search engines: Mathematical modeling and text retrieval", Society for Industrial and Applied Mathematics Philadelphia, 2005.

[38] G.R. Notess, "Search Engine Statistics: Dead Links", Feb. 2000, available at (Sep. 21, 2004): http://www.searchengineshowdown.com/stats/dead.shtml

[39] N. KANDO, NTCIR (NII-NACSIS test collection for IR systems). Project NTCIR Home, 2002, available at (Mar. 9, 2008): http://research.nii.ac.jp/ntcir/.

[40] N. L. Fielden, L. Kuntz, "Search engines handbook", McFarland & Company Inc., 2002.

[41] PC Computing, "Comparing the Top Ten Chinese Search Engine", (in Chinese), available at (Sep. 21, 2004): http://www.net345.com/comnet/sousuo--intro.htm

[42] Popular Computer Week E-version, "A Report on Commonly Used Search Engines", (in Chinese), available at (Sep. 21, 2004): http://www.ahzx.net/frontpage/CHAP3_7_3.HTML

[43] P. Diaconis, R. L. Graham, "Spearman's footrule as a measure of disarray", Journal of the Royal Statistical Society, Series B (Methodological), 39, 1977, pp. 262-268.

[44] R. Fagin, R. Kumar, D. Sivakumar, "Comparing top k lists", SIAM Journal on Discrete Mathematics, 17(1), 2003, pp. 108-118.

[45] R.R. Korfhage, "Information Storage and Retrieval", John Wiley & Sons, Inc., New York, 1997.

[46] S. Brin, L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", Proceedings of the Seventh International World Wide Web Conference, 1998.

[47] S. Chakrabarti, B. Dom, R. S. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, J. M. Kleinberg, D. Gibson, "Hyper searching the Web", Scientific American, 280(6), pp. 54-60, 1999

[48] S. Clarke, P. Willett, "Estimating the recall performance of search engines", ASLIB Proceedings, 49 (7), pp. 184-189, 1997.

[49] S. K. Kwan, "An Economic Model for Comparing Search Services", Proceedings of the 39th Hawaii International Conference on System Sciences, 2006

[50] S. Lawrence, C. L. Giles, "Accessibility of information on the Web", Nature, 400, pp. 107-109, 1999.

[51] L. Si, J. Callan, "Using Sampled Data and Regression to Merge Search Engine Results," ACM SIGIR'02, Aug. 11-15, 2002, Finland.

[52] Shanghai Society for Scientific and Technical Information, "A Research on Chinese Search Engine Comparison", (in Chinese) available at (Mar. 9, 2008): http://www.widewaysearch.com/paper3.htm

[53] T. Agata, M. Nozue, N. Hattori, S. Ueda, "A measure for evaluating search engines on the World Wide Web: Retrieval test with ESL", Library and Information, 37, pp. 1-11, 1997.

[54] Tsinghua University IT Usability Lab, "Search Engine Comparison Report", (in Chinese) available at (Sep. 21, 2004): http://news.ccidnet.com/pub/article/c951_a127264_p1.html

[55] Valencia Community College, "Web Search Engines Comparison", available at (Mar. 9, 2008): http://valencia.cc.fl.us/lrcwest/searchchart.html

[56] Van, "Foundations of evaluation", Journal of Documentation, 30, pp 365–373, 1974.

[57] W. J. Kuo, R. –F. Chang, "Approximating the statistical distribution of color histogram for content-based image retrieval", Proceedings of the Acoustics, Speech, and Signal Processing, Volume 04, 2000, pp. 2007-2010.

[58] W.S. Cooper, "Expected search length: A single measure of retrieval effectiveness based on the weak ordering action of retrieval systems", American Documentation, 19(1), pp. 30–41, 1968.

[59] Y. Aharoni, A. J. Frank, "Finding information on the free World Wide Web: A specialty meta-search engine for the academic community", 2005, available at (Mar. 9, 2008): http://www.firstmonday.org/issues/issue10_12/aharoni/index.html.

[60] Y. Liu, Y. Fu, M. Zhang, S. Ma, L. Ru, "Automatic Search Engine Performance Evaluation with Click-through Data Analysis". WWW 2007 Poster, May 8-12, 2007.

[61] 13th Statistical Survey Report on the Internet Development in China, Feb 2004. China Internet Network Information Centre (CNNIC), available at (Mar. 9, 2008): http://www.cnnic.cn/download/manual/en-reports/13.pdf

[62] 14th Statistical Survey Report on the Internet Development in China, Jul 2004. China Internet Network Information Centre (CNNIC), available at (Mar. 9, 2008): http://www.cnnic.cn/download/2004/2004072003.pdf

[63] 15th Statistical Survey Report on the Internet Development in China, Jan 2005. China Internet Network Information Centre (CNNIC), available at (Mar. 9, 2008): http://www.cnnic.cn/download/2005/2005012701.pdf

[64] 16th Statistical Survey Report on the Internet Development in China, Jul 2005. China Internet Network Information Centre (CNNIC), available at (Mar. 9, 2008): http://www.cnnic.cn/download/2005/2005072601.pdf

[65] 17th Statistical Survey Report on the Internet Development in China, Feb 2006. China Internet Network Information Centre (CNNIC), available at (Mar. 9, 2008): http://www.cnnic.cn/download/2006/17threport-en.pdf

[66] 18th Statistical Survey Report on the Internet Development in China, Aug 2006. China Internet Network Information Centre (CNNIC), available at (Mar. 9, 2008): http://www.cnnic.cn/download/2006/18threport-en.pdf

Appendix

Results using Tsunami as the keyword are shown here for the three performance measurement metrics, for both Chinese and English search engines. For the complete sets of results, please visit: http://www.ece.uvic.ca/~wangyl.

A.1 Daily Duplication Frequency, Chinese Search Engines

Keyword	Parameters	Google	Yahoo	Baidu	Tianwang	Zhongsou
Tsunami From Sep. 12, 2006 to Oct. 12, 2006	Median	0.00	0.00	0.00	0.00	0.00
	Standard Deviation	0.22	0.24	0.21	0.25	0.27
	Skewness	2.80	3.12	2.94	3.21	3.24
	ExpectedValue	10.69	10.72	10.00	10.74	10.90
	Distance from Group DD-R	0.99	0.53	2.06	0.33	0.12
	Distance from Group DD-U	47.19	50.52	41.46	52.03	54.70
	Distance from Group DD-L	94.78	95.01	82.06	95.07	98.05
	Group to which the engine belongs	Group DD-R	Group DD-R	Group DD-R	Group DD-R	Group DD-R

Keyword	Parameters	Google	Yahoo	Baidu	Tianwang	Zhongsou
Tsunami From Oct. 13, 2006 to Nov. 13, 2006	Median	0.00	0.00	0.00	0.00	0.00
	Standard Deviation	0.22	0.22	0.27	0.24	0.20
	Skewness	2.95	2.95	3.25	3.16	2.84
	ExpectedValue	10.68	10.68	10.90	10.69	10.42
	Distance from Group DD-R	0.84	0.84	0.11	0.48	1.51
	Distance from Group DD-U	48.20	48.20	54.95	50.72	44.24
	Distance from Group DD-L	94.39	94.39	98.17	94.33	89.90
	Group to which the engine belongs	Group DD-R	Group DD-R	Group DD-R	Group DD-R	Group DD-R

A.2 Period Duplication Frequency, Chinese Search Engines

Keyword	Parameters	Google	Yahoo	Baidu	Tianwang	Zhongsou
Tsunami From Sep. 12, 2006 to Oct. 12, 2006	Median	0.00	0.00	0.00	0.00	0.00
	Standard Deviation	0.11	0.07	0.07	0.11	0.15
	Skewness	4.50	3.04	3.54	2.39	5.28
	ExpectedValue	22.00	19.06	18.00	10.53	28.09
	Distance from Group PD-R	9.59	21.55	21.92	51.99	0.99
	Distance from Group PD-L	45.59	39.93	33.92	19.14	73.54
	Distance from Group PD-M	4.97	2.03	0.96	4.11	18.54
	Group to which the engine belongs	Group PD-M	Group PD-M	Group PD-M	Group PD-M	Group PD-R

Keyword	Parameters	Google	Yahoo	Baidu	Tianwang	Zhongsou
Tsunami From Oct. 13, 2006 to Nov. 13, 2006	Median	0.00	0.00	0.00	0.00	0.00
	Standard Deviation	0.06	0.09	0.11	0.11	0.06
	Skewness	2.99	4.25	4.96	4.55	2.54
	ExpectedValue	8.89	6.40	5.42	22.00	9.65
	Distance from Group PD-R	25.28	11.74	2.75	9.42	32.66
	Distance from Group PD-L	35.94	43.74	66.75	45.42	32.66
	Distance from Group PD-M	1.47	3.90	14.20	5.04	1.74
	Group to which the engine belongs	Group PD-M	Group PD-M	Group PD-R	Group PD-M	Group PD-M

A.3 Daily Rank Change Frequency, Chinese Search Engines

Keyword	Parameters	Google	Yahoo	Baidu	Tianwang	Zhongsou
Tsunami From Sep. 12, 2006 to Oct. 12, 2006	Median	0.00	0.00	0.00	0.71	0.45
	Standard Deviation	11.91	17.24	20.58	6.57	12.46
	Skewness	3.64	4.27	4.35	3.76	4.27
	ExpectedValue	10.33	10.27	10.66	10.41	10.10
	Distance from Group DR-N	95.55	97.70	91.38	93.38	99.68
	Distance from Group DR-Z	2.17	3.01	4.60	1.59	1.75
	Distance from Group DR-N-Z	23.74	27.23	25.15	22.75	27.62
	Group to which the engine belongs	Group DR-Z	Group DR-Z	Group DR-Z	Group DR-Z	Group DR-Z

Keyword	Parameters	Google	Yahoo	Baidu	Tianwang	Zhongsou
Tsunami From Oct. 13, 2006 to Nov. 13, 2006	Median	0.00	0.00	0.00	0.71	0.45
	Standard Deviation	11.91	17.24	20.58	6.57	12.46
	Skewness	3.64	4.27	4.35	3.76	4.27
	ExpectedValue	10.32	11.29	10.07	10.35	10.59
	Distance from Group DR-N	95.84	78.79	102.81	94.56	90.35
	Distance from Group DR-Z	2.16	4.61	4.16	1.54	2.09
	Distance from Group DR-N-Z	23.88	18.58	30.64	23.31	23.12
	Group to which the engine belongs	Group DR-Z	Group DR-Z	Group DR-Z	Group DR-Z	Group DR-Z

A.4 Daily Duplication Frequency, English Search Engines

Keyword	Parameters	Google	Yahoo	MSN	Lycos	Hotbot	AOL
Tsunami From Sep. 12, 2006 to Oct. 12, 2006	Median	0.00	0.00	0.00	0.00	0.00	0.00
	Standard Deviation	23.39	26.16	20.95	19.07	18.18	24.43
	Skewness	3.00	3.19	2.67	2.04	2.08	3.02
	ExpectedValue	10.74	9.61	9.61	10.48	10.42	10.81
	Distance from Group DD-R	5.50	8.63	6.61	5.43	5.07	5.95
	Distance from Group DD-U	49.61	43.03	36.77	39.29	38.61	50.90
	Distance from Group DD-L	100.34	80.88	78.87	95.10	93.45	102.08
	Group to which the engine belongs	Group DD-R	Group DD-R	Group DD-R	Group DD-R	Group DD-R	Group DD-R

Keyword	Parameters	Google	Yahoo	MSN	Lycos	Hotbot	AOL
Tsunami From Oct. 13, 2006 to Nov. 13, 2006	Median	0.00	0.00	0.00	0.00	0.00	0.00
	Standard Deviation	27.00	29.10	18.07	24.05	19.93	28.08
	Skewness	3.25	3.31	2.42	3.17	2.62	3.29
	ExpectedValue	10.90	9.97	10.35	10.55	10.35	10.94
	Distance from Group DD-R	7.14	9.36	4.37	5.86	4.76	7.72
	Distance from Group DD-U	54.83	48.24	40.00	49.42	41.94	56.11
	Distance from Group DD-L	105.14	88.71	91.47	96.83	91.85	106.43
	Group to which the engine belongs	Group DD-R	Group DD-R	Group DD-R	Group DD-R	Group DD-R	Group DD-R

A.5 Period Duplication Frequency, English Search Engines

Keyword	Parameters	Google	Yahoo	MSN	Lycos	Hotbot	AOL
Tsunami From Sep. 12, 2006 to Oct. 12, 2006	Median	0.00	0.00	0.00	0.00	0.00	0.00
	Standard Deviation	0.11	0.11	0.08	0.08	0.08	0.11
	Skewness	4.91	3.63	3.59	4.37	4.03	5.10
	ExpectedValue	24.62	24.00	23.75	22.64	21.13	24.62
	Distance from Group PD-R	5.05	9.54	10.39	9.59	13.40	4.82
	Distance from Group PD-L	56.74	57.54	56.89	49.45	44.20	56.52
	Distance from Group PD-M	9.57	7.28	7.00	5.63	3.45	10.04
	Group to which the engine belongs	Group PD-R	Group PD-M	Group PD-M	Group PD-M	Group PD-M	Group PD-R

Keyword	Parameters	Google	Yahoo	MSN	Lycos	Hotbot	AOL
Tsunami From Oct. 13, 2006 to Nov. 13, 2006	Median	0.00	0.00	0.00	0.00	0.00	0.00
	Standard Deviation	0.14	0.14	0.07	0.10	0.08	0.15
	Skewness	5.28	5.41	3.66	3.92	3.03	5.46
	ExpectedValue	25.83	28.80	19.50	16.74	15.14	29.09
	Distance from Group PD-R	2.98	0.66	18.25	24.02	33.09	0.50
	Distance from Group PD-L	61.98	77.46	39.25	28.45	27.95	79.05
	Distance from Group PD-M	12.91	20.33	1.85	0.24	0.83	21.29
	Group to which the engine belongs	Group PD-R	Group PD-R	Group PD-M	Group PD-M	Group PD-M	Group PD-R

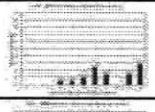
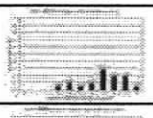
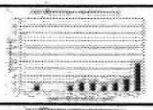
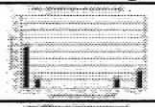
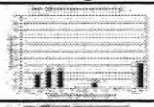
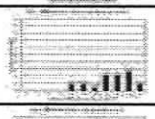
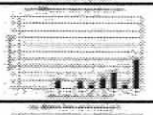
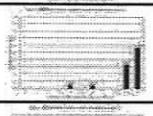
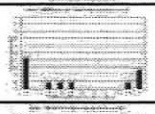
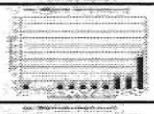
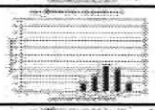
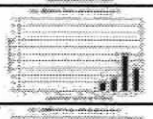
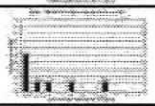
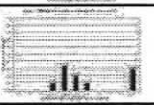
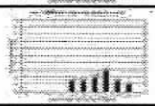
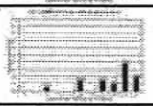
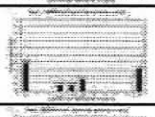
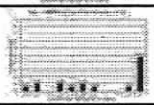
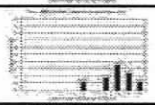
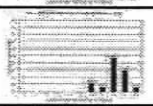
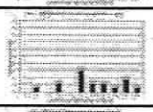
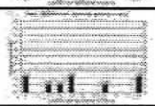
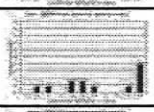
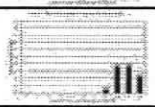
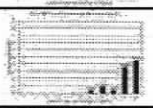
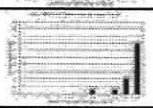
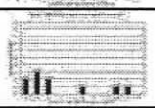
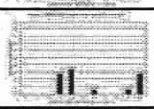
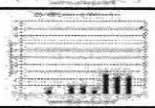
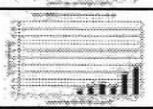
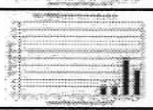
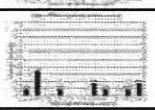
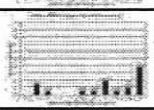
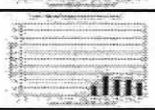
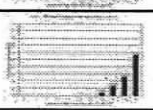
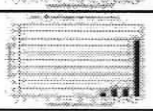
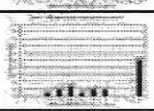
A.6 Daily Rank Chang Frequency, English Search Engines

Keyword	Parameters	Google	Yahoo	MSN	Lycos	Hotbot	AOL
Tsunami From Sep. 12, 2006 to Oct. 12, 2006	Median	0.00	0.00	0.00	0.00	0.00	0.00
	Standard Deviation	20.86	20.70	19.44	17.71	17.56	19.39
	Skewness	4.46	4.46	4.43	4.33	4.32	4.42
	ExpectedValue	10.26	10.43	10.37	10.47	10.53	10.19
	Distance from Group DR-N	99.17	95.80	96.58	95.97	94.91	100.18
	Distance from Group DR-Z	4.34	4.40	4.04	5.28	5.52	3.99
	Distance from Group DR-N-Z	29.25	27.56	27.49	25.72	25.06	29.18
	Group to which the engine belongs	Group DR-Z	Group DR-Z	Group DR-Z	Group DR-Z	Group DR-Z	Group DR-Z

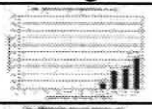
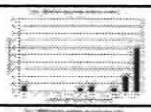
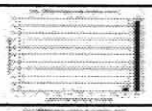
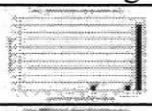
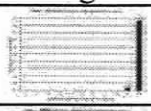
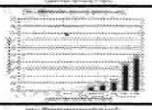
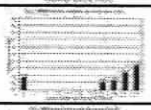
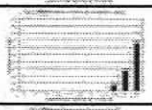
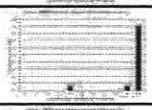
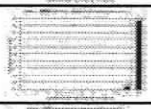
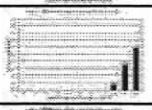
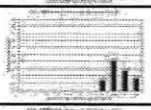
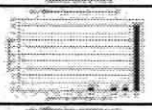
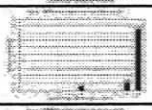
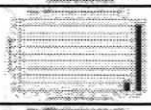
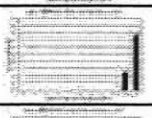
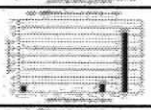
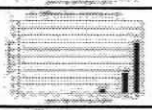
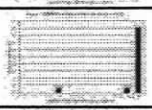
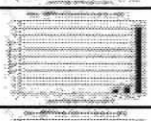
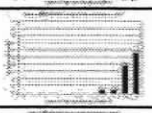
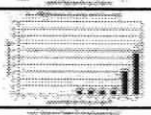
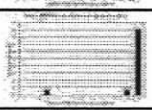
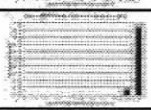
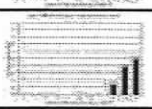
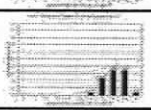
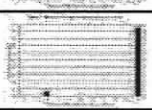
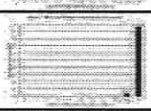
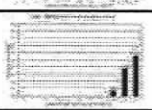
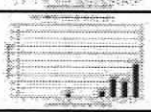
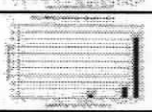
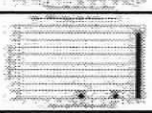
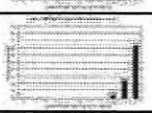
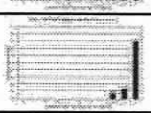
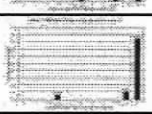
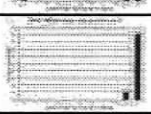
Keyword	Parameters	Google	Yahoo	MSN	Lycos	Hotbot	AOL
Tsunami From Oct. 13, 2006 to Nov. 13, 2006	Median	0.00	0.00	0.00	0.00	0.00	0.00
	Standard Deviation	19.44	22.28	17.71	20.63	20.12	20.64
	Skewness	4.43	4.47	4.33	4.45	4.44	4.45
	ExpectedValue	10.10	10.04	10.37	10.41	10.60	10.07
	Distance from Group DR-N	101.81	104.15	97.84	96.24	92.37	102.85
	Distance from Group DR-Z	3.88	4.86	5.25	4.36	4.45	4.21
	Distance from Group DR-N-Z	30.07	32.05	26.59	27.75	25.71	30.98
	Group to which the engine belongs	Group DR-Z	Group DR-Z	Group DR-Z	Group DR-Z	Group DR-Z	Group DR-Z

A.7 Histograms for Daily Duplication Frequency, Chinese Search Engines

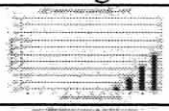
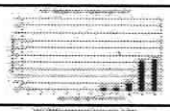
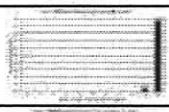
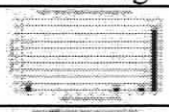
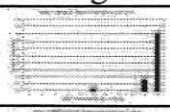
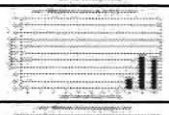
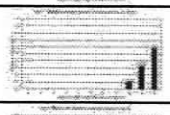
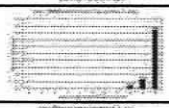
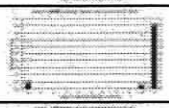
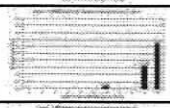
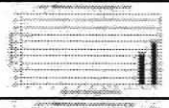
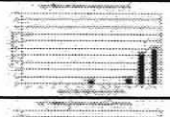
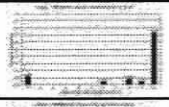
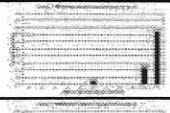
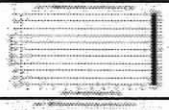
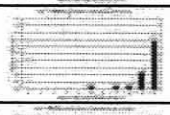
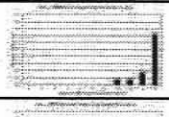
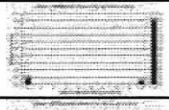
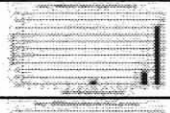
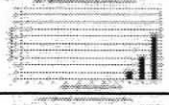
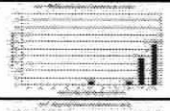
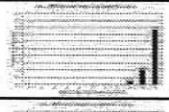
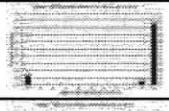
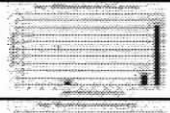
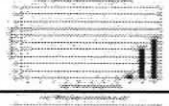
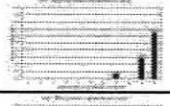
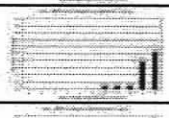
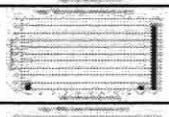
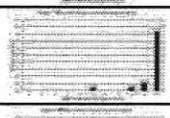
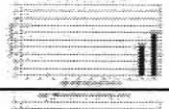
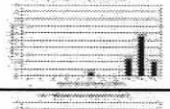
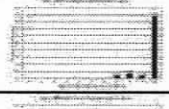
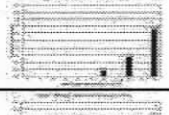
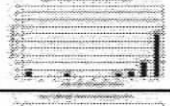
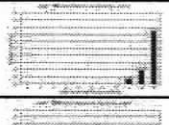
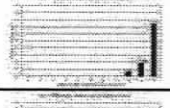
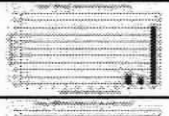
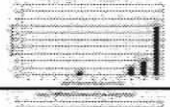
From Sep. 20, 2005 to Oct. 25, 2005

Keyword	Google	Yisou	Baidu	Tianwang	Zhongsou
Chao Nv					
Furong Sister					
Hurricane					
Katrina					
New Orleans					
Oil Price					
Rita					
Tsunami					

From Sep. 12, 2006 to Oct. 12, 2006

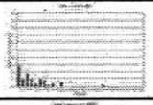
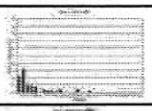
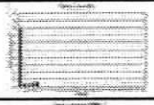
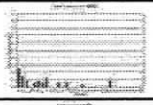
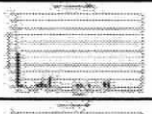
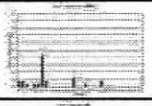
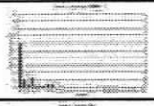
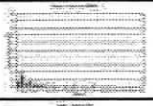
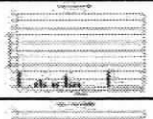
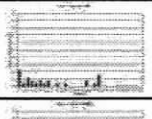
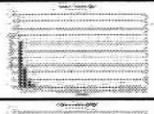
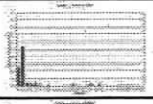
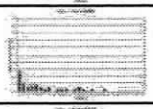
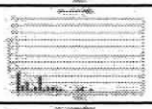
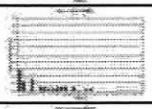
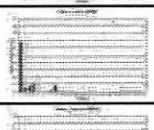
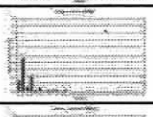
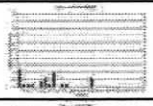
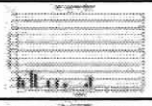
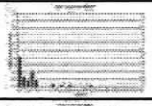
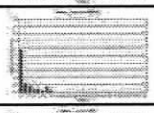
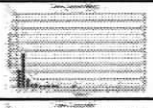
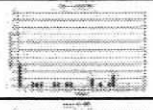
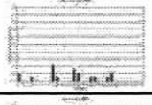
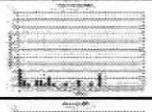
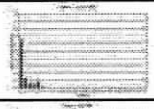
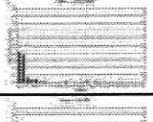
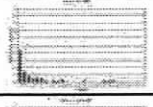
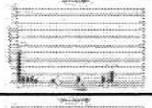
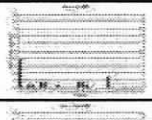
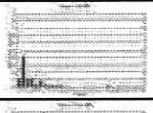
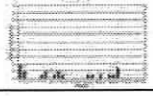
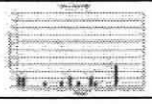
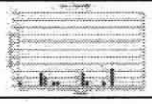
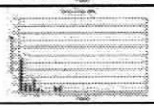
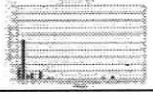
Keyword	Google	Yisou	Baidu	Tianwang	Zhongsou
Chao Nv					
Furong Sister					
Hurricane					
Katrina					
New Orleans					
Oil Price					
Rita					
Tsunami					
Bird Flu					
Hu Jintao					
Chen Shuibian					
WTO					

From Oct. 13, 2006 to Nov. 13, 2006

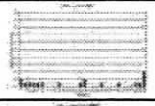
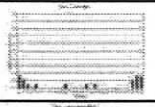
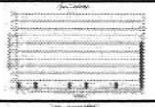
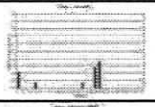
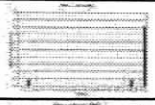
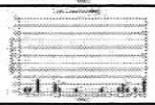
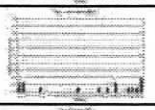
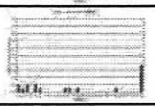
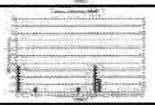
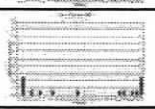
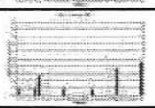
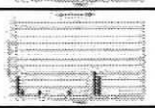
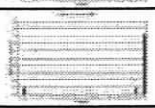
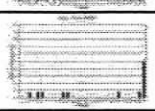
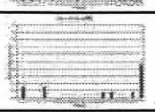
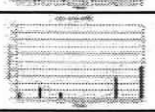
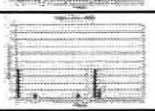
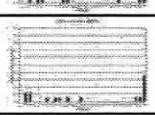
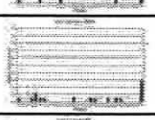
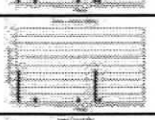
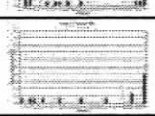
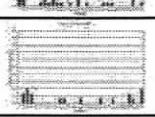
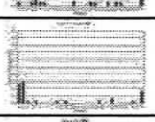
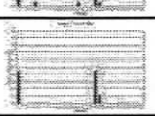
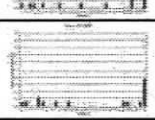
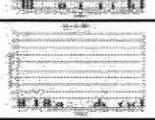
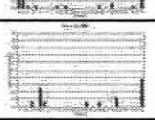
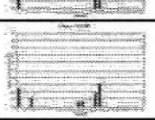
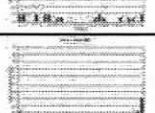
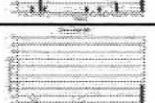
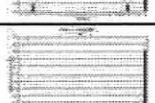
Keyword	Google	Yisou	Baidu	Tianwang	Zhongsou
Chao Nv					
Furong Sister					
Hurricane					
Katrina					
New Orleans					
Oil Price					
Rita					
Tsunami					
Bird Flu					
Hu Jintao					
Chen Shuibian					
WTO					

A.8 Histograms for Period Duplication Frequency, Chinese Search Engines

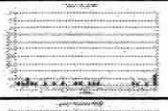
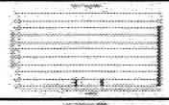
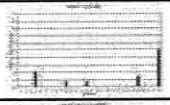
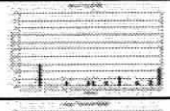
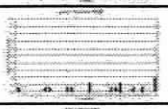
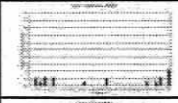
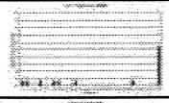
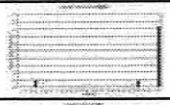
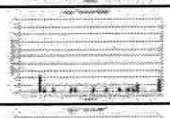
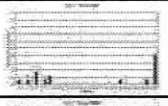
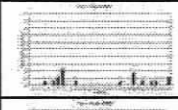
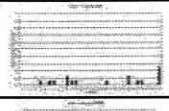
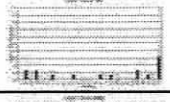
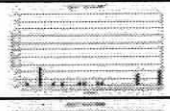
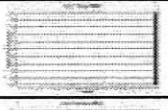
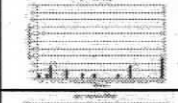
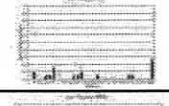
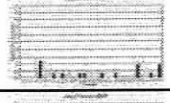
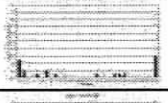
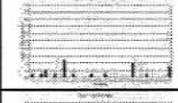
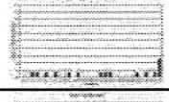
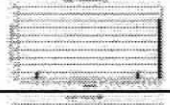
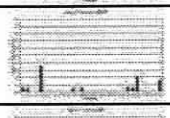
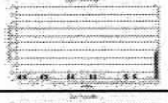
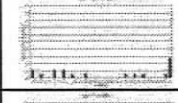
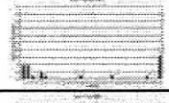
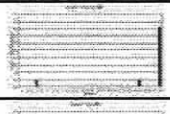
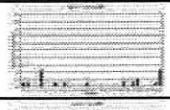
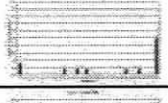
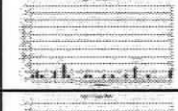
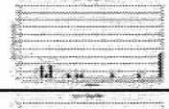
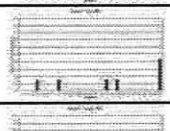
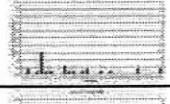
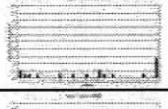
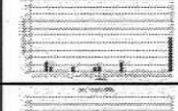
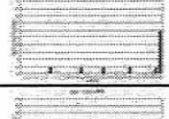
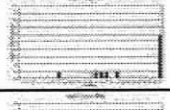
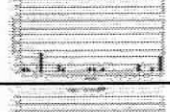
From Sep. 20, 2005 to Oct. 25, 2005

Keyword	Google	Yisou	Baidu	Tianwang	Zhongsou
Chao Nv					
Furong Sister					
Hurricane					
Katrina					
New Orleans					
Oil Price					
Rita					
Tsunami					

From Sep. 12, 2006 to Oct. 12, 2006

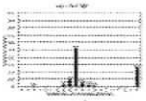
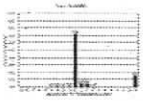
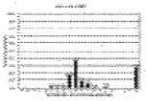
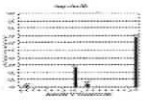
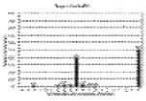
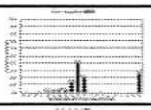
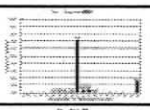
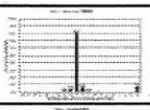
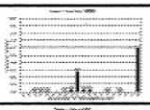
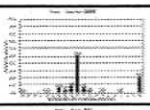
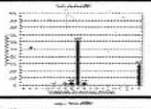
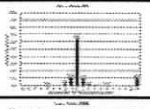
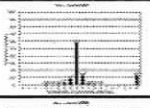
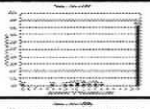
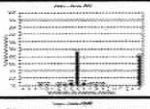
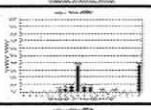
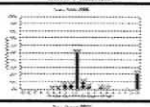
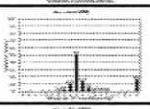
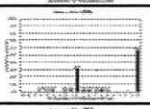
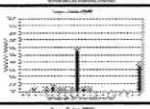
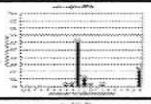
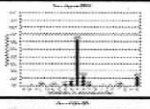
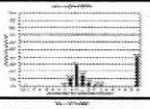
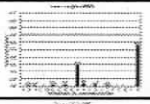
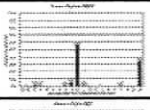
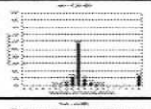
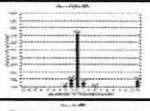
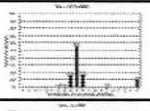
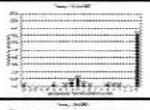
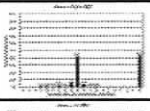
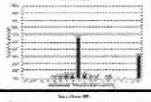
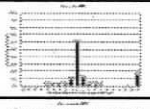
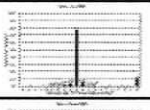
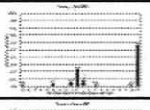
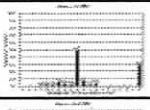
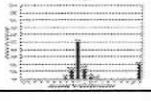
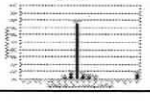
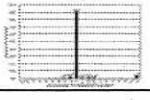
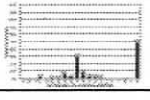
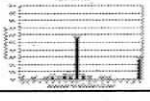
Keyword	Google	Yisou	Baidu	Tianwang	Zhongsou
Chao Nv					
Furong Sister					
Hurricane					
Katrina					
New Orleans					
Oil Price					
Rita					
Tsunami					
Bird Flu					
Hu Jintao					
Chen Shuibian					
WTO					

From Oct. 13, 2006 to Nov. 13, 2006

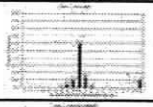
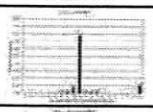
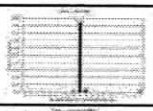
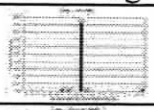
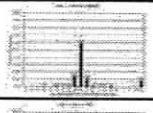
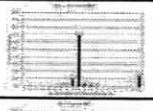
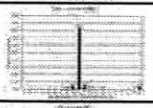
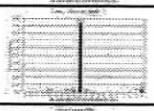
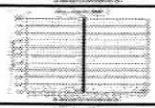
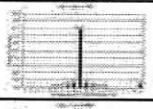
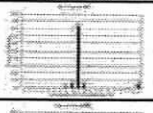
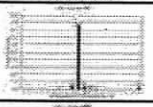
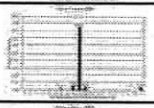
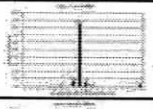
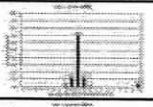
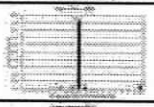
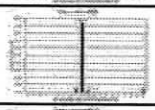
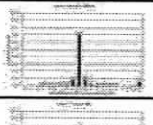
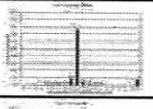
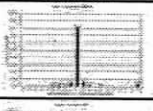
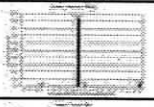
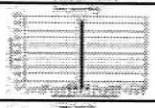
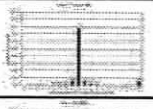
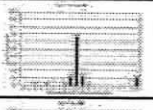
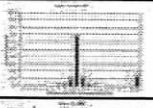
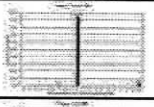
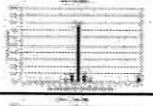
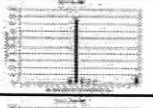
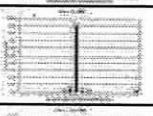
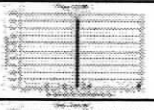
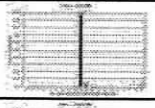
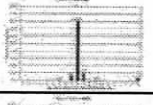
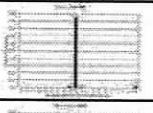
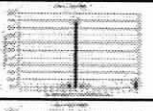
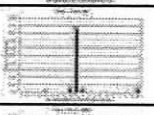
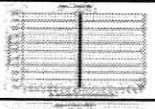
Keyword	Google	Yisou	Baidu	Tianwang	Zhongsou
Chao Nv					
Furong Sister					
Hurricane					
Katrina					
New Orleans					
Oil Price					
Rita					
Tsunami					
Bird Flu					
Hu Jintao					
Chen Shuibian					
WTO					

A.9 Histograms for Daily Rank Change Frequency, Chinese Search Engines

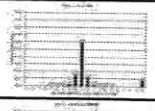
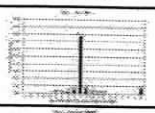
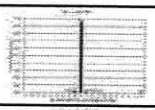
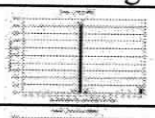
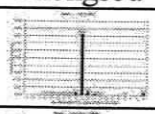
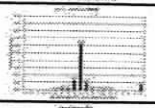
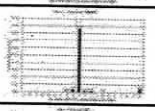
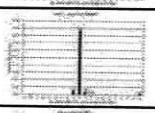
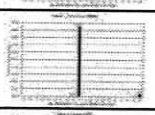
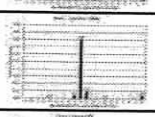
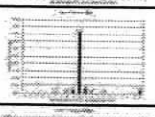
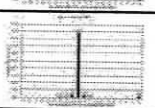
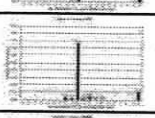
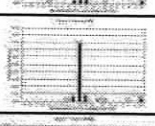
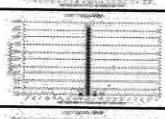
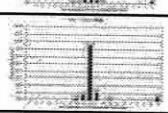
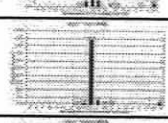
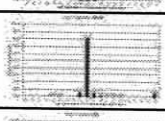
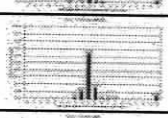
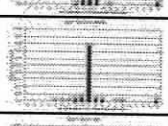
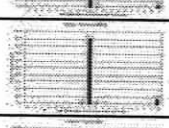
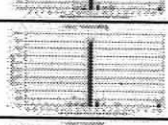
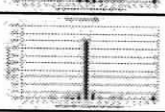
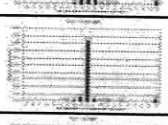
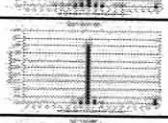
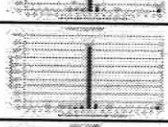
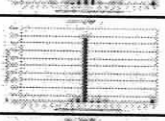
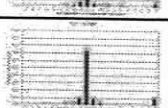
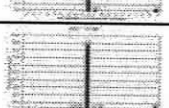
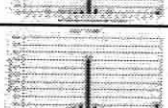
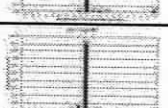
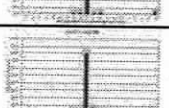
From Sep. 20, 2005 to Oct. 25, 2005

Keyword	Google	Yisou	Baidu	Tianwang	Zhongsou
Chao Nv					
Furong Sister					
Hurricane					
Katrina					
New Orleans					
Oil Price					
Rita					
Tsunami					

From Sep. 12, 2006 to Oct. 12, 2006

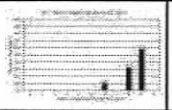
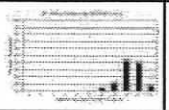
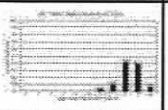
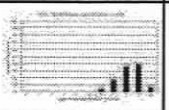
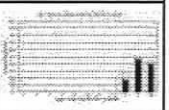
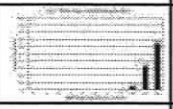
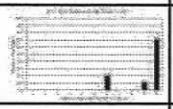
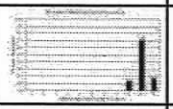
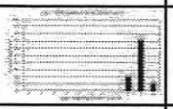
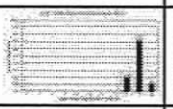
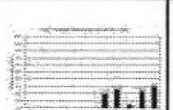
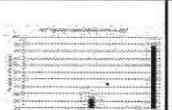
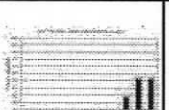
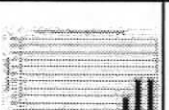
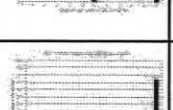
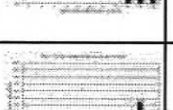
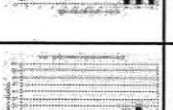
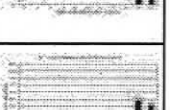
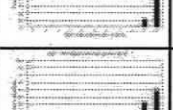
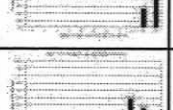
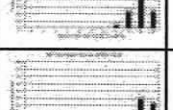
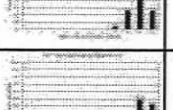
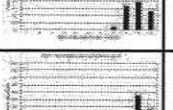
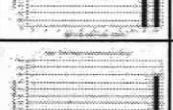
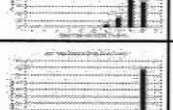
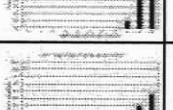
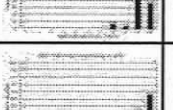
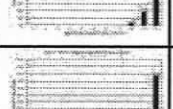
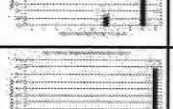
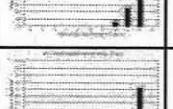
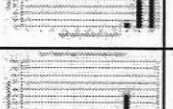
Keyword	Google	Yisou	Baidu	Tianwang	Zhongsou
Chao Nv					
Furong Sister					
Hurricane					
Katrina					
New Orleans					
Oil Price					
Rita					
Tsunami					
Bird Flu					
Hu Jintao					
Chen Shuibian					
WTO					

From Oct. 13, 2006 to Nov. 13, 2006

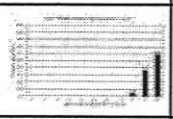
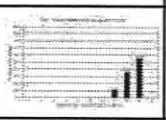
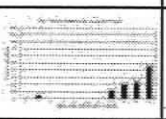
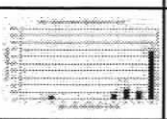
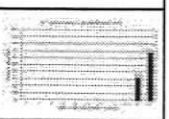
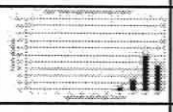
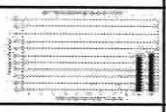
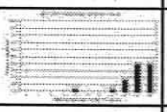
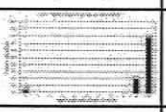
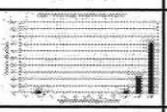
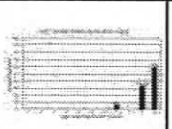
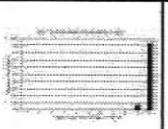
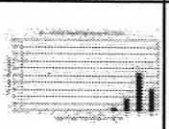
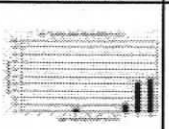
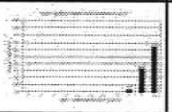
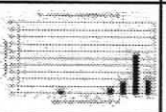
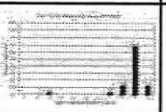
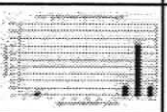
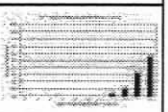
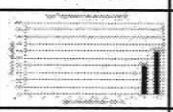
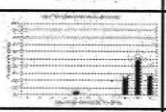
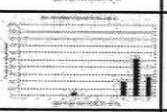
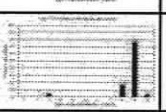
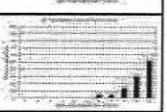
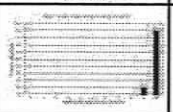
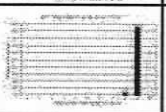
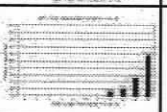
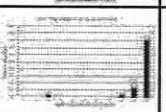
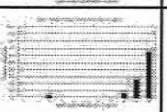
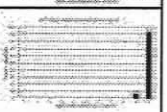
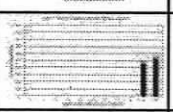
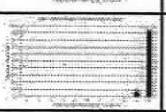
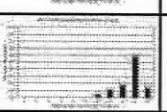
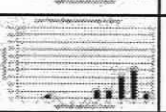
Keyword	Google	Yisou	Baidu	Tianwang	Zhongsou
Chao Nv					
Furong Sister					
Hurricane					
Katrina					
New Orleans					
Oil Price					
Rita					
Tsunami					
Bird Flu					
Hu Jintao					
Chen Shuibian					
WTO					

A.10 Histograms for Daily Duplication Frequency, English Search Engines

From Sep. 12, 2006 to Oct. 12, 2006

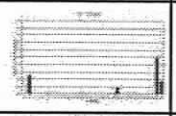
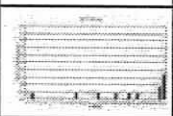
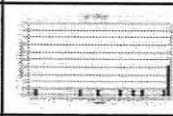
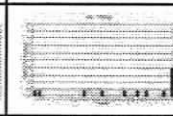
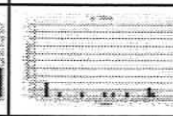
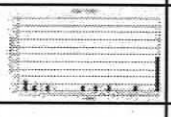
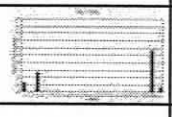
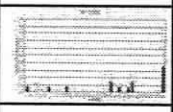
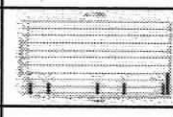
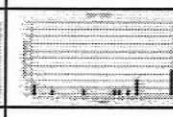
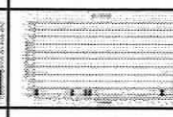
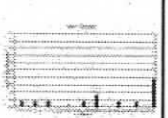
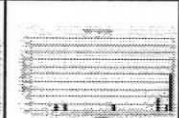
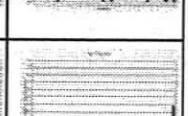
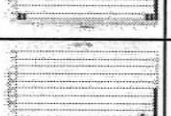
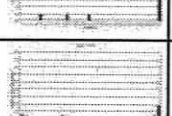
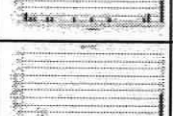
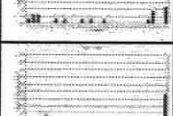
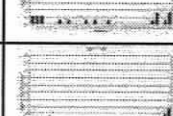
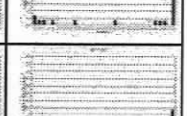
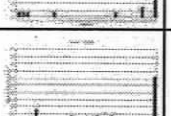
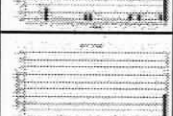
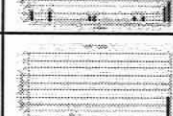
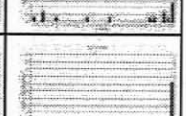
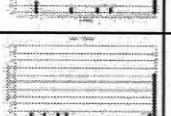
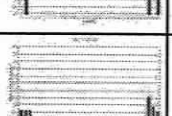
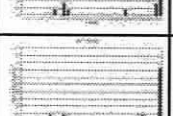
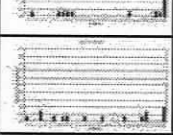
Keyword	Google	Yahoo	MSN	Lycos	Hotbot	AOL
Hurricane						
Katrina						
New Orleans						
Oil Price						
Rita						
Tsunami						
Bird Flu						
Hu Jintao						
Chen Shuibian						
WTO						

From Oct. 13, 2006 to Nov. 13, 2006

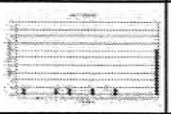
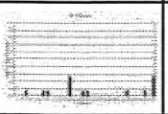
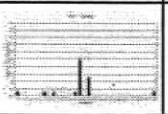
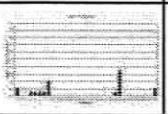
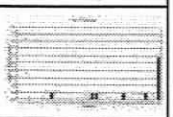
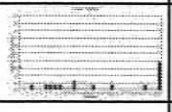
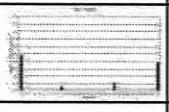
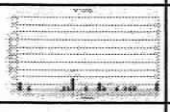
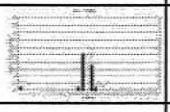
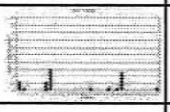
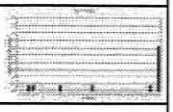
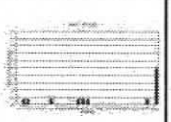
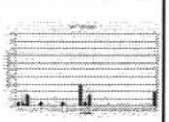
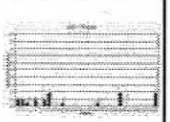
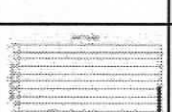
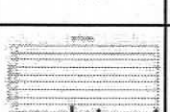
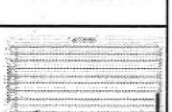
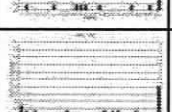
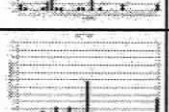
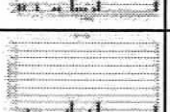
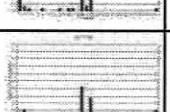
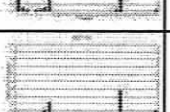
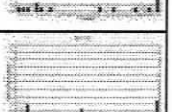
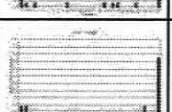
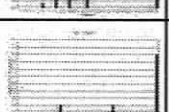
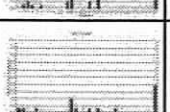
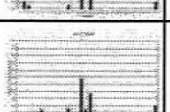
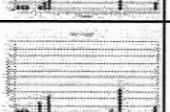
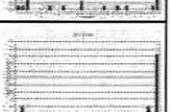
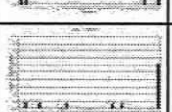
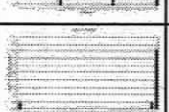
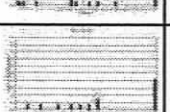
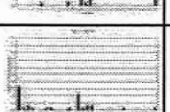
Keyword	Google	Yahoo	MSN	Lycos	Hotbot	AOL
Hurricane						
Katrina						
New Orleans						
Oil Price						
Rita						
Tsunami						
Bird Flu						
Hu Jintao						
Chen Shuibian						
WTO						

A.11 Histograms for Period Duplication Frequency, English Search Engines

From Sep. 12, 2006 to Oct. 12, 2006

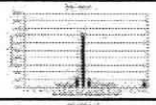
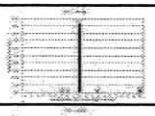
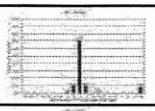
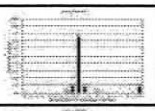
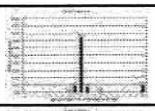
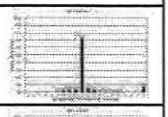
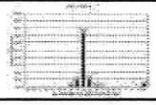
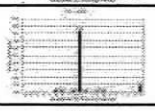
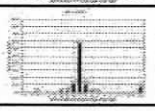
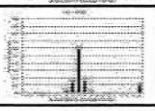
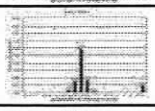
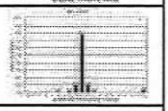
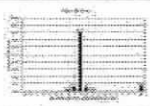
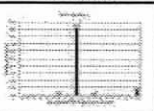
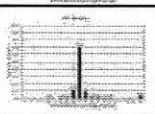
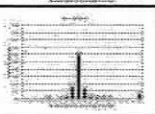
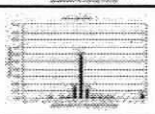
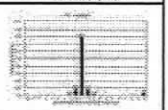
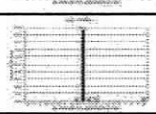
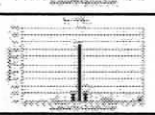
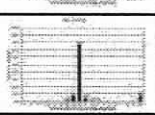
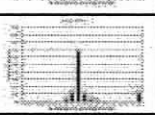
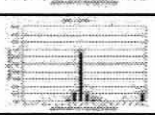
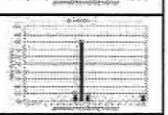
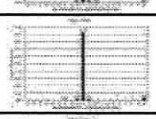
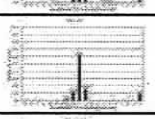
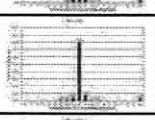
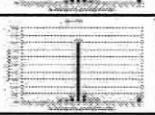
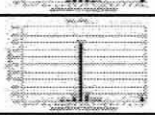
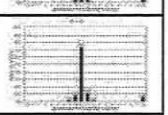
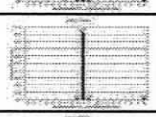
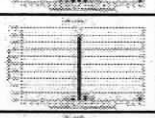
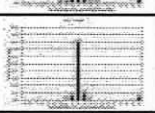
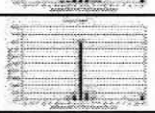
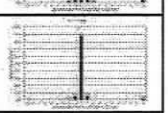
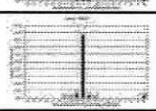
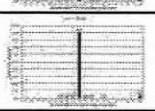
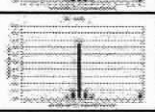
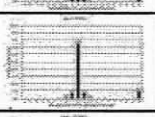
Keyword	Google	Yahoo	MSN	Lycos	Hotbot	AOL
Hurricane						
Katrina						
New Orleans						
Oil Price						
Rita						
Tsunami						
Bird Flu						
Hu Jintao						
Chen Shuibian						
WTO						

From Oct. 13, 2006 to Nov. 13, 2006

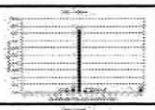
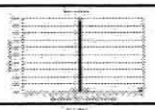
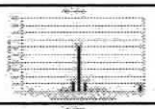
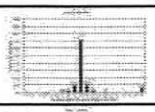
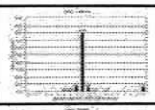
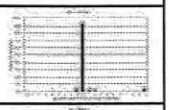
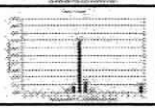
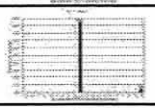
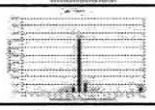
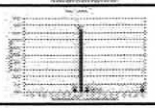
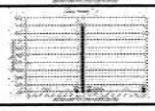
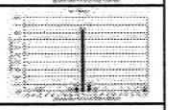
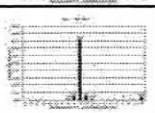
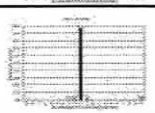
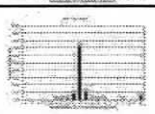
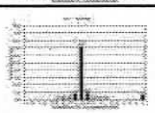
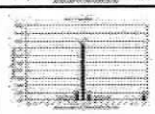
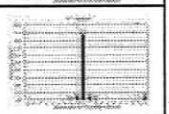
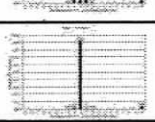
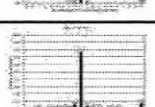
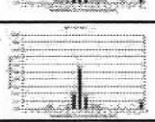
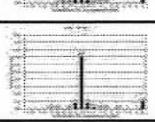
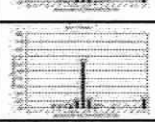
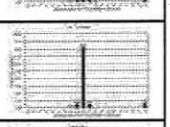
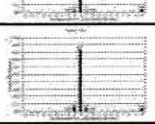
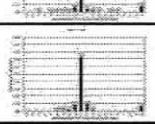
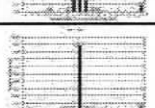
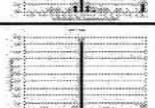
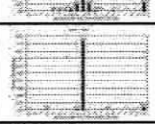
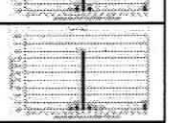
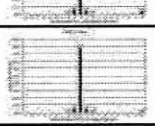
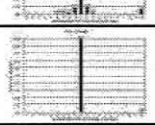
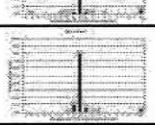
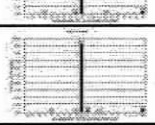
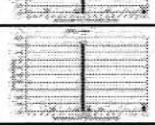
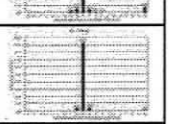
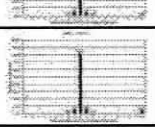
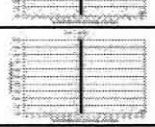
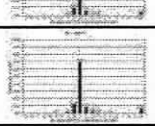
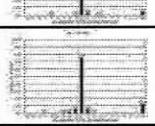
Keyword	Google	Yahoo	MSN	Lycos	Hotbot	AOL
Hurricane						
Katrina						
New Orleans						
Oil Price						
Rita						
Tsunami						
Bird Flu						
Hu Jintao						
Chen Shuibian						
WTO						

A.12 Histograms for Daily Rank Change Frequency, English Search Engines

From Sep. 12, 2006 to Oct. 12, 2006

Key word	Google	Yahoo	MSN	Lycos	Hotbot	AOL
Hurricane						
Katrina						
New Orleans						
Oil Price						
Rita						
Tsunami						
Bird Flu						
Hu Jintao						
Chen Shuibian						
WTO						

From Oct. 13, 2006 to Nov. 13, 2006

Keyword	Google	Yahoo	MSN	Lycos	Hotbot	AOL
Hurricane						
Katrina						
New Orleans						
Oil Price						
Rita						
Tsunami						
Bird Flu						
Hu Jintao						
Chen Shuibian						
WTO						