ENHANCING RELATIONAL DATABASES WITH SEMANTIC SEARCH USING WORD EMBEDDINGS

GERRY PENG, DEPARTMENT OF COMPUTER SCIENCE, MARCH 2026

SEARCH DATA BY MEANING INSTEAD OF KEYWORDS

INTRODUCTION

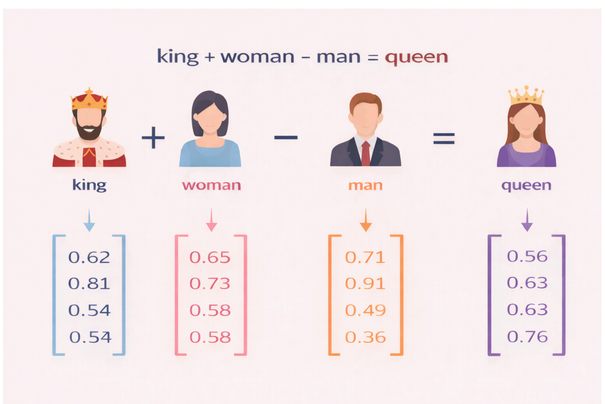
- Relational databases are well suited for categorical data because they support efficient filtering, indexing, and relational joins.
- However, text retrieval in SQL systems relies primarily on exact or partial string matching operations (e.g., LIKE queries).
- These methods are sensitive to wording differences and fail to capture semantic relationships between conceptually similar terms.
- As a result, relevant entries will not be retrieved when synonyms or paraphrasing are used.
- Word embeddings represent words as vectors that encode semantic relationships geometrically.

Research Goal: Enable meaning-aware search in relational databases by integrating word embeddings into the retrieval process.

```
SELECT *
FROM Movies
WHERE title LIKE '%crime%'
```

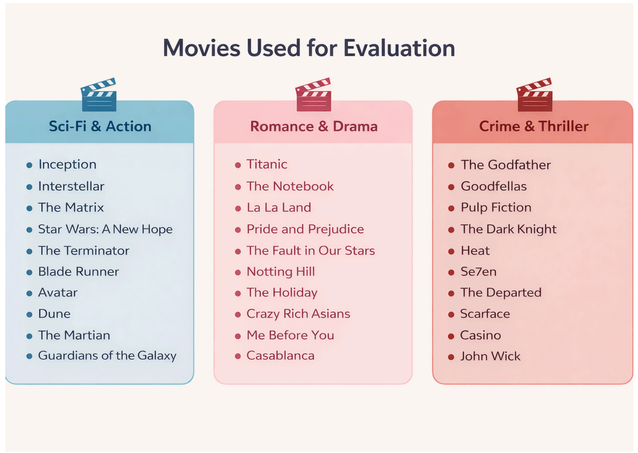
WORD EMBEDDINGS & SEMANTIC RELATIONSHIPS

- Word embeddings map tokens (words) to high-dimensional vector representations learned from large text corpora.
- Words that appear in similar contexts occupy nearby regions in the embedding space.
- Cosine similarity measures angular proximity between vectors, allowing semantic similarity to be quantified numerically.
- Embedding spaces have geometric properties in which semantic relationships can be expressed through vector arithmetic.
- For example, king + woman − man ≈ queen illustrates how relational attributes are encoded algebraically.
- The behavior of embeddings depends on the diversity and scope of their training data.



TEXT REPRESENTATION

- A cloud-hosted MySQL database was constructed containing 30 films across three genres: Sci-Fi/Action, Romance/Drama, and Crime/Thriller.
- Each movie entry included a title and short plot description stored within a relational schema.
- To enable semantic comparison, each film was represented by averaging pretrained GloVe embeddings from its title and plot text.
- The resulting vector served as a compact representation of the movie's meaning.

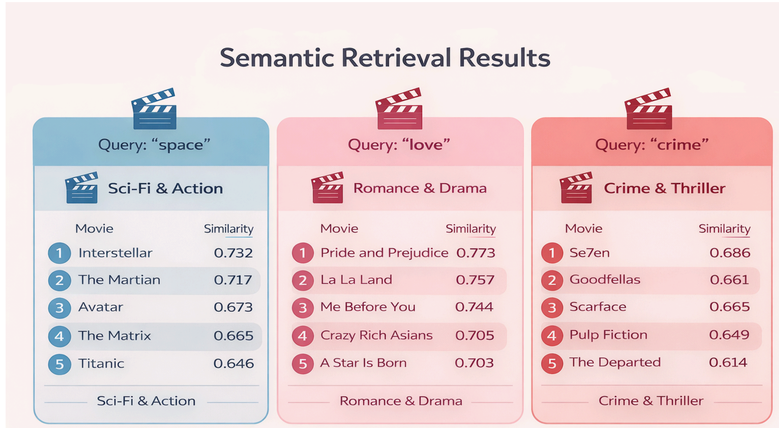


EMBEDDING-BASED SEMANTIC SEARCH

- A user-provided query word is first mapped to its corresponding embedding vector.
- Cosine similarity is computed between the query vector and each movie vector.
- Movies are ranked according to similarity score, producing an ordered list of conceptually related movies.
- Unlike keyword search, this method retrieves thematically relevant movies even when the query word does not explicitly appear in the text.
- The framework transforms text retrieval into geometric similarity comparison in vector space.

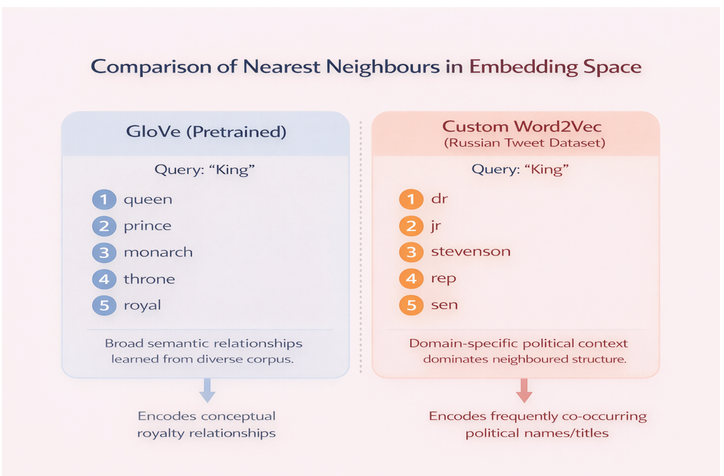
$$\begin{bmatrix} W_{11} \\ W_{12} \\ \\ W_{1n} \end{bmatrix} + \begin{bmatrix} W_{21} \\ W_{22} \\ \\ W_{2n} \end{bmatrix} + \cdots + \begin{bmatrix} W_{n1} \\ W_{n2} \\ \\ W_{nn} \end{bmatrix} = \begin{bmatrix} \frac{W_{11}+W_{21}+\ldots+W_{n1}}{n} \\ \\ \\ \frac{W_{1n}+W_{2n}+\ldots+W_{nn}}{n} \end{bmatrix} D$$

SEMANTIC RETRIEVAL RESULTS



IMPACT OF TRAINING DATA

- Pretrained GloVe embeddings were compared with a custom Word2Vec model trained on a 1GB dataset of 2016 American election-related tweets.
- The custom model captured geopolitically relevant relationships strongly due to domain-specific training.
- However, general vocabulary terms did not perform as well.
- These findings illustrate that embedding representations reflect the properties of their training corpus.
- The effectiveness of semantic retrieval systems therefore depends on the diversity and alignment of the embedding dataset.



This research was supported by the Jamie Cassels Undergraduate Research Awards, University of Victoria. Supervised by Dr Sean Chester