Table of Contents

Report

Introduction

[TBC]

Literature Review

Following paragraphs are the literature review around the chosen topic.

Full-Text Search Engine

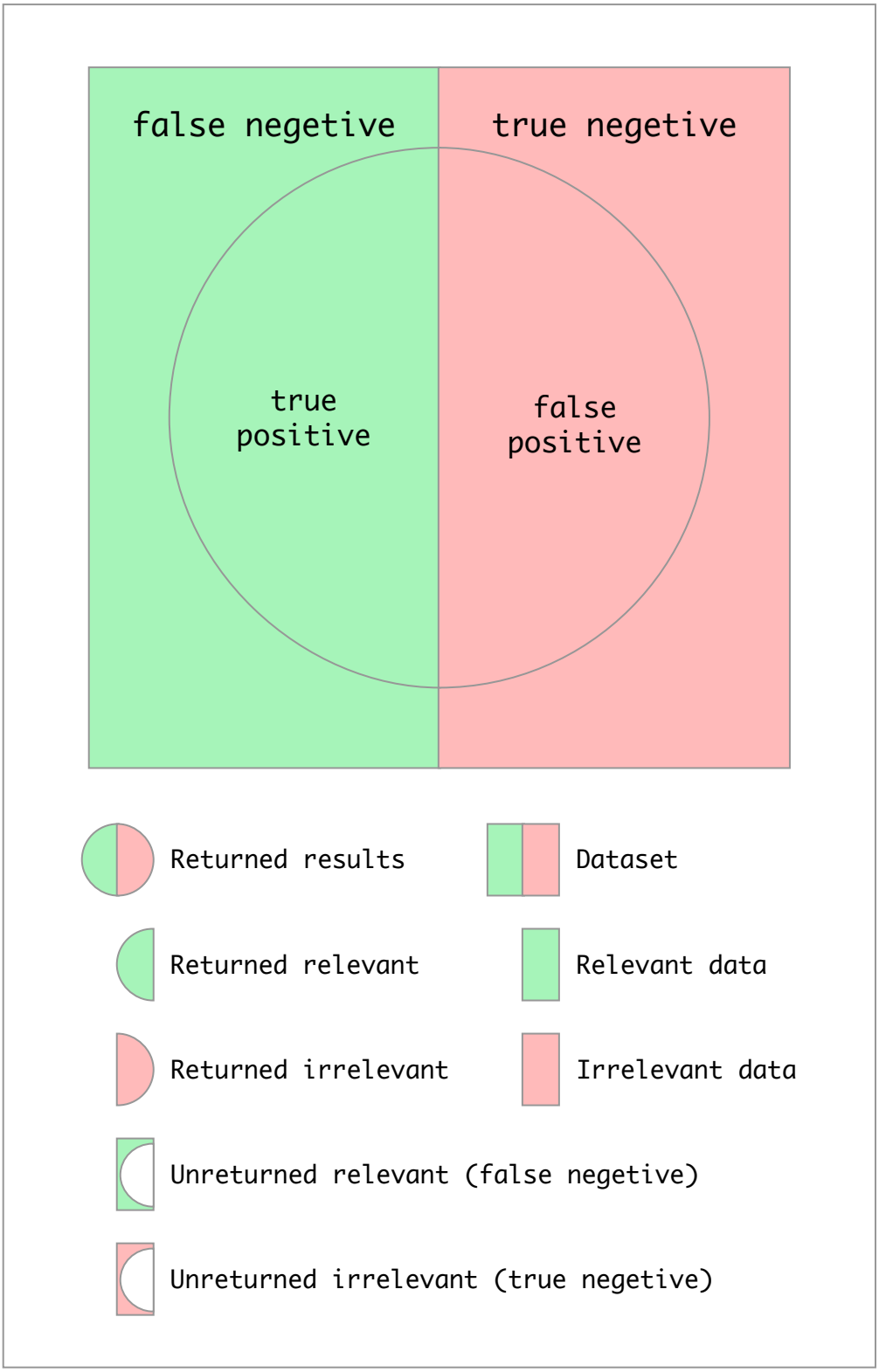
Full-Text search engines perform linguistic searches against text data in target document or collection of text data based on the rules of a particular language (Microsoft Docs, 2020). Full-Text search engines consider the basic unit of Full-Text search as a token rather than a word (Melton & Buxton, 2006). Full-Text queries can include simple words and phrases or multiple forms of a word or phrase. A Full-Text query returns any documents that contain at least one match (also known as a hit). A match occurs when a target document contains all the terms specified in the Full-Text query, and meets any other search conditions, such as the distance between the matching terms. It widely used in modern technology like web search(Google Search, Bing Search, etc.), desktop search(search for local files) and even email search.

As the need of semantic-aware Full-text search and indexing against databases and documents keeps increasing in the past decade, full-text search systems became available to no-expert users who aren't familiar with documents they are searching, don't know the exact vocabulary used to index relevant documents, or don't know how to formulated advanced search queries (Tekli, Chbeir, et al., 2018). All of those issues mentioned above result in noisy or irrelevant search results and false positives & false negatives in the search result (See Precision and Recall).

Precision and Recall

Precision and recall are often used to measure the performance of information retrieval or extraction systems, precision (aka. positive prediction value) is the fraction of relevant results among all results returned (in other words, how many returned results are relevant?), while recall (aka. sensitivity) is the fraction of the amount of

returned relevant results among all relevant results in the dataset (in other words, how many relevant results are returned?) (Makhoul, Kubala, et al., 1999).



In general, there is a tradeoff between "precision" and "recall". High precision means that fewer irrelevant results are returned as a result of the query (fewer false positives), while high recall means that fewer relevant results are missing (fewer false negatives). traditional text-matching systems such as the LIKE operator in SQL gives you 100% precision with no concessions for recall. A full text search facility gives you a lot of flexibility to tune down the precision for better recall (Erickson, 2008).

For example, the SQL LIKE operator can be extremely inefficient. If you apply it to an un-indexed column, a full scan will be used to find matches (just like any query on an un-indexed field). If the column is indexed,

matching can be performed against index keys, but with far less efficiency than most index lookups. Therefore, most full text search implementations use an "inverted index" system to improve performance (see Inverted Index) (Erickson, 2008).

Inverted Index

Inverted Index is an index where the keys are individual terms, and the associated values are sets of records that contain the term. Full text search is optimized to compute the intersection, union, etc. of these record sets, and usually provides a ranking algorithm to quantify how strongly a given record matches search keywords (Erickson, 2008).

furthermore SemIndex (Tekli, Chbeir, et al., 2018)

problem with false positives and false negatives

Morphological Analysis (Stemming)

[TBC]

Ranking

[TBC]

References

Erickson, 2008. sql - What is Full Text Search vs LIKE - Stack Overflow. Accessed on July 1st, 2020.

Melton, J. Buxton, S., 2006. Chapter 13 - What's Missing? - Querying XML | ScienceDirect. Accessed on July 1st, 2020.

Microsoft Docs, 2018. Full-Text Search - SQL Server | Microsoft Docs. Accessed on July 1st, 2020.

Reiner, K. Chichao, C. Farzin, M. Ravi, K.,2006. Searching with context | ACM Digital Library. Accessed on July 1st, 2020.

Tekli, J., Chbeir, R., Traina, A., Traina, C Jr., Yetongnon, K., Ibanez, C., Assad, M., Kallas, C., 2018. Full-fledged semantic indexing and querying model designed for seamless integration in legacy RDBMS. Accessed on July 2nd, 2020.

Makhoul, J., Kubala, F., Schwartz, R. and Weischedel, R., 1999, February. Performance measures for information extraction. In Proceedings of DARPA broadcast news workshop (pp. 249-252).