Use of artificial intelligence in a field of obtaining information from documents using probabilistic model*

Alžbeta Žiarovská
Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
xziarovska@stuba.sk

26. september 2023

*Semestral project in subject Engineering Methods, ac. year 2023/24, guidance: MSc. Mirwais Ahmadzai

Abstract

. . .

1 Introduction

This article concludes creations of many scientists and researchers who were all devoted to finding out as much as possible in a field of information retrieval. "Information retrieval" is a enormously broad term. I would like to work with a focus on document retrieval, meaning, extracting information from documents written using natural language.

There are many methods, which we can use to retrieve information, which will be closer addressed in subsection "Other types of models" of section "Models" 3.2 However, the main focus of this article is going to be probabilistic model, which is more deeply explained in subsection "Probabilistic model" of section "Models" 3.1.

Artificial intelligence has become everyday part of lives of many people. We are using a great variety of tools to help us find appropriate information we can use for numerous purposes (e.g. education, medicine,...) and, apparently, we are getting lazier by doing so. [AHA+23]

Spoločenské súvisloti Toto je paragraf, neviem, ako sa ukáže, to idem teraz zistiť, aka fuka funda luka... Čo sa stane? Rozbijem to? Pokračujem v Introduction, alebo? Možno ani nepokračujem [Jon99]a možno pokračujem.

2 Information retrieval from documents

Information retrieval from documents (further in article I will refer to this term only as document retrieval) is a main term of this paper. It is referring to a primarily linguistic process of extracting information from textual material or documents. The least we have to do in this process is to describe want we are yearning for and make this description compatible with descriptions of information we have access to. Furthermore, our description of what we are looking for must have a meaning. [Bla03]

3 Models of Information Retrieval

By using terminology an information retrieval system (model) we understand a software algorithm which stores and manages information obtained from documents (often textual, but multimedia is also a possibility). The purpose of the system is to assist a user in finding the exact information they are looking for and need. The system does not explicitly answer questions or return information. However, it offers a information that contains existence and location of documents which might possibly hold the information desired by the user. The main goal is to satisfy user's need of information and some of the found documents will hopefully achieve user's satisfaction. The documents that fulfill this goal are called *relevant documents*. A retrieval system that is perfect would only retrieve documents that are relevant and no irrelevant documents would be offered. Nevertheless, such system does not exist and will never exist either, mostly because relevance is a subjective opinion of the user.

We distinguish three primal processes an information retrieval system has to be able to provide:

1. The representation of the content of the documents
2. The representation of the user's information need

3. The comparison of the two representations

These processes are shown in the Figure 1. In the figure, blue colored boxes represent processed and red colored boxes represent data.

Indexing usually refers to process of representing the documents. The process happens off-line, meaning, there is no direct involvement of the end user of the information retrieval system. The result of the indexing process is a representation of the document. Documents are often only stored partially, for example only the title and the abstract, plus the actual location of the document. However the indexing might include the actual storage of the document.

The process of representing the user's *need for information* if often called the *query formulation process*. In general, query formulation may denote the complete interactive dialogue between user and system. This leads not to only accurate query but possibly also to bettering user's understanding of their information needs. This part is visualized by the feedback process in Figure 1.

The comparison of the query against the document representations is referred to as *matching process*. The process ordinarily results in a ranked list of documents. User is able to go through the list and search for the information they need. To minimize the time user needs to find the information, ranked retrieval hopefully puts the relevant documents on the top of the list. Effective ranking algorithms are rather simple, they use frequency distribution of terms over documents as well as statistics over other information. Effective ranking algorithms might easily halve the time user spends on reading documents. [Hie09]

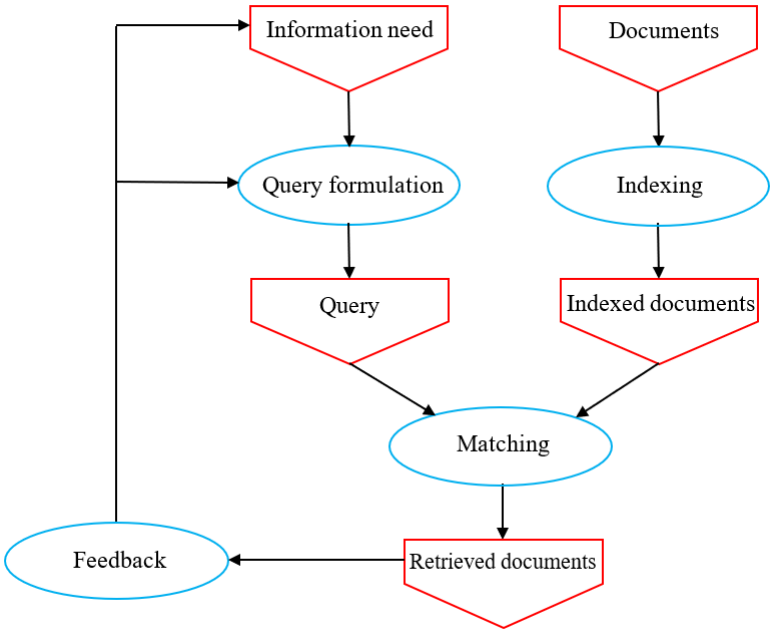

Figure 1: Information Retrieval Processes

3.1 Probabilistic model

[Hie09] [Jon99]

3.2 Other types of models

There will be two information retrieval models mentioned and addressed in this section. Both of them provide exact matching (documents are either retrieved or not), but there is no ranking of the retrieved documents.

The Boolean model The Boolean model is believed to be the first information retrieval model and probably is also one of the most criticized ones. It can be explained by defining query as a unambiguous definition of a set of written documents. To give an example the query term economic defines the group of all documents that are indexed with the label economic. New sets of documents can be creating by combining query terms and their competent sets of documents using the operators of Boole's mathematical logic.

The main advantage of this type of IR model is giving a sense of control over the system to the user. A user can be certain, why has the document been retrieved. According to the document sets' size it is immediately clear which operators of Boolean algebra (AND, OR, NOT) will produce a smaller or bigger set of documents. We can see a visualized example in a Figure 2[Hie09]

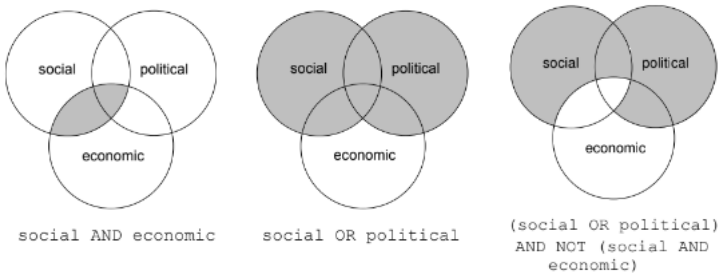

Figure 2: Visualization of Boolean combinations of sets by Venn diagrams

The vector space model The main point of approach by vector space model is to model information retrieval objects as elements of vector space. To be more specific, terms, documents, queries and so on are all vectors in the vector space. By having the vector space we have a system with linear properties, for example the ability to add together two elements to get a new one or the ability to multiply a vector by a real number.The similarity between documents and queries can be measured using these properties, such as the scalar product of the corresponding vectors.

In a vector space model, the key components are:

- *Dimension*: The number of terms and keywords that determines the size of the vectors representing documents and queries.

- *Basis Vectors*: A set of vectors that form a basis from the vector space. In some cases, these basis vectors may be assumed to be pairwise orthogonal, although this is not always realistic.
- *Correlations*: If the term vectors are not orthogonal, their correlations can be used to represent the relationships between terms.
- *Document vectors*: These vectors represent the documents in the vector space. The components of these vectors correspond to the weights of the terms within the documents.
- *Query vectors*: Represent the user queries, their components correspond to the weights of the terms in the queries.
- *Similarity measurement*: A method of determining the similarity between two vectors, such as the scalar product. This similarity can be used to rank documents in the order of their estimated usefulness to a user.

[RW86]

4 Conclusion

Z obr. 3 je všetko jasné.

Aj text môže byť prezentovaný ako obrázok. Stane sa z neho označný plávajúci objekt. Po vytvorení diagramu zrušte znak `%` pred príkazom `\includegraphics` označte tento riadok ako komentár (tiež pomocou znaku `%`).

Figure 3: Rozhodujuci argument.

5 Iná časť

Základným problémom je teda. . . Najprv sa pozrieme na nejaké vysvetlenie (časť 6.1), a potom na ešte nejaké (časť 6.1).[1]

Môže sa zdať, že problém vlastne nejestvuje[Cop99], ale bolo dokázané, že to tak nie je [CHE05, CK05]. Napriek tomu, aj dnes na webe narazíme na všelijaké pochybné názory[SEI]. Dôležité veci možno *zdôrazniť kurzívou*.

6 Ďalšia časť

Toto je ďalšia časť, v ktorej idem urobiť odsek.

Toto je odsek. haha.

[1] Niekedy môžete potrebovať aj poznámku pod čiarou.

6.1 Nejaké vysvetlenie

Niekedy treba uviesť zoznam:

- jedna vec
- druhá vec
  - x
  - y

Ten istý zoznam, len číslovaný:

1. jedna vec
2. druhá vec
   (a) x
   (b) y

6.2 Ešte nejaké vysvetlenie

Veľmi dôležitá poznámka. Niekedy je potrebné nadpisom označiť odsek. Text pokračuje hneď za nadpisom.

7 Dôležitá časť

8 Ešte dôležitejšia časť

9 Záver

References

[AHA+23] Sayed Fayaz Ahmad, Heesup Han, Muhammad Mansoor Alam, Mohd Rehmat, Muhammad Irshad, Marcelo Arraño-Muñoz, Antonio Ariza-Montes, et al. Impact of artificial intelligence on human loss in decision making, laziness and safety in education. *Humanities and Social Sciences Communications*, 10(1):1–14, 2023.

[Bla03] David C Blair. Information retrieval and the philosophy of language. 2003.

[CHE05] Krzysztof Czarnecki, Simon Helsen, and Ulrich Eisenecker. Staged configuration through specialization and multi-level configuration of feature models. *Software Process: Improvement and Practice*, 10:143–169, April/June 2005.

[CK05] Krzysztof Czarnecki and Chang Hwan Peter Kim. Cardinality-based feature modeling and constraints: A progress report. In *International Workshop on Software Factories, OOPSLA 2005*, San Diego, USA, October 2005.

[Cop99] James O. Coplien. *Multi-Paradigm Design for C++*. Addison-Wesley, 1999.

[Hie09] Djoerd Hiemstra. Information retrieval models. *Information Retrieval: searching in the 21st Century*, pages 1–19, 2009.

[Jon99] Karen Sparck Jones. Information retrieval and artificial intelligence. *Artificial Intelligence*, 114(1-2):257–281, 1999.

[RW86] Vijay V Raghavan and SK Michael Wong. A critical analysis of vector space model for information retrieval. *Journal of the American Society for information Science*, 37(5):279–287, 1986.

[SEI] Carnegie Mellon University Software Engineering Institute. A framework for software product line practice—version 5.0. `http://www.sei.cmu.edu/productlines/frame_report/`.