CT4100

Information Retrieval

Name: Andrew Hayes
E-mail: a.hayes18@universityofgalway.ie
Student ID: 21321503

2024–09–19

Contents

1 Introduction

1.1 Lecturer Contact Details

- Colm O'Riordan.
- `colm.oriordan@universityofgalway.ie`.

1.2 Motivations

- To study/analyse techniques to deal suitably with the large amounts (& types) of information.
- Emphasis on research & practice in Information Retrieval.

1.3 Related Fields

- Artificial Intelligence.
- Database & Information Systems.
- Algorithms.
- Human-Computer Interaction.

1.4 Recommended Texts

- *Modern Information Retrieval* – Riberio-Neto & Baeza-Yates (several copies in library).
- *Information Retrieval* – Grossman.
- *Introduction to Information Retrieval* – Christopher Manning.
- Extra resources such as research papers will be recommended as extra reading.

1.5 Grading

- Exam: 70%.
- Assignment 1: 30%.
- Assignment 2: 30%.

There will be exercise sheets posted for most lecturers; these are not mandatory and are intended as a study aid.

1.6 Introduction to Information Retrieval

Information Retrieval (IR) deals with identifying relevant information based on users' information needs, e.g. web search engines, digital libraries, & recommender systems. It is finding material (usually documents) of an unstructured nature that satisfies an information need within large collections (usually stored on computers).

2 Information Retrieval Models

2.1 Introduction to Information Retrieval Models

Data collections are well-structured collections of related items; items are usually atomic with a well-defined interpretation. Data retrieval involves the selection of a fixed set of data based on a well-defined query (e.g., SQL, OQL).

Information collections are usually semi-structured or unstructured. Information Retrieval (IR) involves the retrieval of documents of natural language which is typically not structured and may be semantically ambiguous.

2.1.1 Information Retrieval vs Information Filtering

The main differences between information retrieval & information filtering are:

- The nature of the information need.
- The nature of the document set.

Other than these two differences, the same models are used. Documents & queries are represented using the same set of techniques and similar comparison algorithms are also used.

2.1.2 User Role

In traditional IR, the user role was reasonably well-defined in that a user:

- Formulated a query.
- Viewed the results.
- Potentially offered feedback.
- Potentially reformulated their query and repeated steps.

In more recent systems, with the increasing popularity of the hypertext paradigm, users usually intersperse browsing with the traditional querying. This raises many new difficulties & challenges.

2.2 Pre-Processing

Document pre-processing is the application of a set of well-known techniques to the documents & queries prior to any comparison. This includes, among others:

- **Stemming:** the reduction of words to a potentially common root. The most common stemming algorithms are Lovin's & Porter's algorithms. E.g. *computerisation*, *computing*, *computers* could all be stemmed to the common form *comput*.
- **Stop-word removal:** the removal of very frequent terms from documents, which add little to the semantics of meaning of the document.
- **Thesaurus construction:** the manual or automatic creation of thesauri used to try to identify synonyms within the documents.

Representation & comparison technique depends on the information retrieval model chosen. The choice of feedback techniques is also dependent on the model chosen.

2.3 Models

Retrieval models can be broadly categorised as:

- Boolean:
  - Classical Boolean.
  - Fuzzy Set approach.
  - Extended Boolean.
- Vector:
  - Vector Space approach.
  - Latent Semantic indexing.
  - Neural Networks.

- Probabilistic:
  - Inference Network.
  - Belief Network.

We can view any IR model as being comprised of:

- D is the set of logical representations within the documents.
- Q is the set of logical representations of the user information needs (queries).
- F is a framework for modelling representations (D & Q) and the relationship between D & Q.
- R is a ranking function which defines an ordering among the documents with regard to any query q.

We have a set of index terms:

$$t_1, \ldots, t_n$$

A **weight** $w_{i,j}$ is assigned to each term t_i occurring in the d_j. We can view a document or query as a vector of weights:

$$\vec{d_j} = (w_1, w_2, w_3, \ldots)$$

2.4 Boolean Model

The **Boolean model** of information retrieval is based on set theory & Boolean algebra. A query is viewed as a Boolean expression. The model also assumes terms are present or absent, hence term weights $w_{i,j}$ are binary & discrete, i.e., $w_{i,j}$ is an element of $\{0, 1\}$.

Advantages of the Boolean model include:

- Clean formalism.
- Widespread & popular.
- Relatively simple

Disadvantages of the Boolean model include:

- People often have difficulty formulating expressions, harbours some difficulty in use.
- Documents are considered either relevant or irrelevant; no partial matching allowed.
- Poor performance.
- Suffers badly from natural language effects of synonymy etc.
- No ranking of results.
- Terms in a document are considered independent of each other.

2.4.1 Example

$$q = t_1 \wedge (t_2 \vee (\neg t_3))$$

```
1 q = t1 AND (t2 OR (NOT t3))
```

This can be mapped to what is termed **disjunctive normal form**, where we have a series of disjunctions (or logical ORs) of conjunctions.

$$q = 100 \vee 110 \vee 111$$

If a document satisfies any of the components, the document is deemed relevant and returned.

2.5 Vector Space Model

The **vector space model** attempts to improve upon the Boolean model by removing the limitation of binary weights for index terms. Terms can have non-binary weights in both queries & documents. Hence, we can represent the documents & the query as n-dimensional vectors.

$$\vec{d_j} = (w_{1,j}, w_{2,j}, \ldots, w_{n,j})$$
$$\vec{q} = (w_{1,q}, w_{2,q}, \ldots, w_{n,q})$$

We can calculate the similarity between a document & a query by calculating the similarity between the vector representations of the document & query by measuring the cosine of the angle between the two vectors.

$$\vec{a} \cdot \vec{b} = \mid \vec{a} \mid\mid \vec{b} \mid \cos(\vec{a}, \vec{b})$$

$$\Rightarrow \cos(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{\mid \vec{a} \mid\mid \vec{b} \mid}$$

We can therefore calculate the similarity between a document and a query as:

$$\text{sim}(q, d) = \cos(\vec{q}, \vec{d}) = \frac{\vec{q} \cdot \vec{d}}{\mid \vec{q} \mid\mid \vec{d} \mid}$$

Considering term weights on the query and documents, we can calculate similarity between the document & query as:

$$\text{sim}(q, d) = \frac{\sum_{i=1}^{N}(w_{i,q} \times w_{i,d})}{\sqrt{\sum_{i=1}^{N}(w_{i,q})^2} \times \sqrt{\sum_{i=1}^{N}(w_{i,d})^2}}$$

Advantages of the vector space model over the Boolean model include:

- Improved performance due to weighting schemes.
- Partial matching is allowed which gives a natural ranking.

The primary disadvantage of the vector space model is that terms are considered to be mutually independent.

2.5.1 Weighting Schemes

We need a means to calculate the term weights in the document and query vector representations. A term's frequency within a document quantifies how well a term describes a document; the more frequently a term occurs in a document, the better it is at describing that document and vice-versa. This frequency is known as the **term frequency** or **tf factor**.

If a term occurs frequently across all the documents, that term does little to distinguish one document from another. This factor is known as the **inverse document frequency** or **idf-frequency**. Traditionally, the most commonly-used weighting schemes are know as **tf-idf** weighting schemes.

For all terms in a document, the weight assigned can be calculated as:

$$w_{i,j} = f_{i,j} \times \log\left(\frac{N}{N_i}\right)$$

where

- $f_{i,j}$ is the (possibly normalised) frequency of term t_i in document d_j.
- N is the number of documents in the collection.
- N_i is the number of documents that contain term t_i.