THE ANNALS OF "DUNAREA DE JOS" UNIVERSITY OF GALATI
FASCICLE III, 2007 ISSN 1221-454X
ELECTROTECHNICS, ELECTRONICS, AUTOMATIC CONTROL, INFORMATICS

VECTOR SPACE MODEL FOR DOCUMENT REPRESENTATION IN INFORMATION RETRIEVAL

Dan MUNTEANU

"Dunărea de Jos" University of Galatz
Faculty of Computer Science
Department of Computers and Applied Informatics
111 Domnească Street, 800201-Galatz, Romania
Phone/Fax: (+40) 236 460182; (+40) 236 461353
E-mail: dan.munteanu@ugal.ro

Abstract: this paper presents the basics of information retrieval: the vector space model for document representation with Boolean and term weighted models, ranking methods based on the cosine factor and evaluation measures: recall, precision and combined measure.

Keywords: information retrieval, Boolean vector space model, term weighted vector space model, document representation

1. INTRODUCTION

Text is the primary way that human knowledge is stored, and after speech, the primary way it is transmitted. Techniques for storing and searching for textual documents are nearly as old as written language itself. Computing, however, has changed the ways text is stored, searched, and retrieved. In traditional library indexing, for example, documents could only be accessed by a small number of index terms such as title, author, and a few subject headings. With automated systems, the number of indexing terms that can be used for an item is virtually limitless.

The subfield of computer science that deals with the automated storage and retrieval of documents is called information retrieval (IR). Automated IR systems were originally developed to help manage the huge scientific literature that has developed since the 1940s, and this is still the most common use of IR systems. IR systems are in widespread use in university, corporate, and public libraries. IR techniques have also been found useful, however, in such disparate areas as office automation and software engineering. Indeed, any field that relies on documents to do its work could potentially benefit from IR techniques (Frakes and Baeza-Yates, 1992).

Information retrieval is an activity, and like most activities it has a purpose. A user of a search engine begins with an information need, which he or she realizes as a query in order to find relevant documents. This query may not be the best articulation of that need, or the best bait to use in a particular document pool. It may contain misspelled, misused, or poorly selected words. It may contain too many words or not enough. Nevertheless, it is usually the only clue that the search engine has concerning the user's goal.

We often speak of documents in the result set as being more or less relevant to the query, but, strictly speaking, this is inaccurate. The user will judge relevance with respect to the information need, not the query. If irrelevant documents are returned, the user may or may not realize why this

This paper was recommended for publication by Severin Bumbaru

is the case, and may or may not find ways to improve the query. The relationship between the query and the documents is explained entirely by the logic of the search engine (Jackson and Moulinier, 2002).

A definition of Information Retrieval (Manning, Raghavan and Schütze, 2007) can be: "Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfy an information need from within large collections (usually on local computer servers or on the internet)".

2. PROBLEM FORMULATION

Most IR systems have as their primary data structure an inverted index. An inverted index is a data structure that lists for each word in the collection all documents that contain it (the postings) and the frequency of occurrence in each document. An inverted index makes it easy to search for hits of a query word. One just goes to the part of the inverted index that corresponds to the query word and retrieves the documents listed there.

A more sophisticated version of the inverted index also contains position information. Instead of just listing the documents that a word occurs in, the positions of all occurrences in the document are also listed.

In some IR systems, not all words are represented in the inverted index. A stop list of function words lists those words that are deemed unlikely to be useful for searching. Common stop words are the, from and could. These words have important semantic functions in English, but they rarely contribute information if the search criterion is a simple word-by-word match. A stop list has the advantage that it reduces the size of the inverted index.

However, it is impossible to search for phrases that contain stop words once the stop list has been applied. For this reason, many retrieval engines do not make use of a stop list for indexing.

Another common feature of IR systems is stemming. In IR, stemming usually refers to a simplified form of morphological analysis consisting simply of truncating a word. For example, laughing, laugh, laughs and laughed are all stemmed to laugh- (Manning and Schütze, 1999).

Relational databases have precise schema and are queried using SQL (structured query language). In relational algebra, the response to a query is always an unordered set of qualifying tuples. Keyword queries are not precise, in the sense that a Boolean decision to include or exclude a response is unacceptable. A safer bet is to rate each document for how likely it is to satisfy the user's information need, sort in decreasing order of this score, and present the results in a ranked list.

The next figure (figure 1) shows the activities associated with a typical IR system.

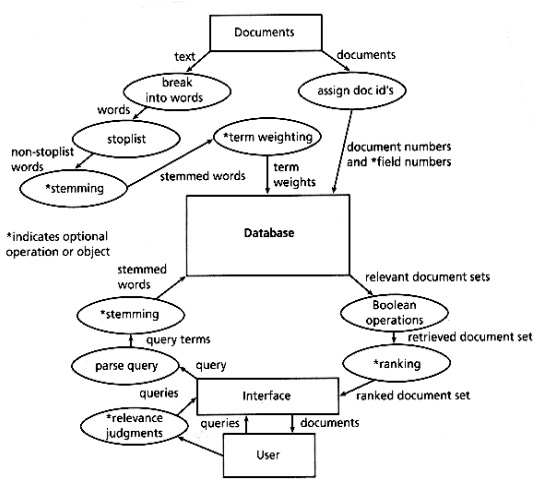

Figure 1. Activities associated with standard IR systems.

Since only a part of the user's information need is expressed through the query, there can be no algorithmic way of ensuring that the ranking strategy always favors the information need. However, mature practice in IR has evolved a vector-space model for documents and a broad class of ranking algorithms based on this model (Chakrabarti, 2003).

A query to IR search engine may return several thousands of matching documents, but a typical user will only be able to examine a small fraction of these. Ranking matching documents according to their relevance to the user is therefore a fundamental problem. In the next sections some classic approaches will be reviewed.

3. VECTOR SPACE MODEL FOR INFORMATION RETRIEVAL

3.1. Boolean Vector Space Model

Text documents can be conveniently represented in a high-dimensional vector space where terms are associated with vector components. More precisely, a text document d can be represented as a sequence of terms, $d = (\omega(1), \omega(2), ..., \omega(|d|))$, where $|d|$ is the length of the document and $\omega(t) \in V$. A vector-space representation of d is

then defined as a real vector $x \in R^{|V|}$, where each component x_j is a statistic related to the occurrence of the j th vocabulary entry in the document. The simplest vector-based representation is Boolean, i.e. $x_j \in \{0,1\}$ indicates the presence or the absence of term ω_j in the document being represented. In the figure 2 is presented a vector-space documentation in the Boolean model and in the Term-Weighted model.

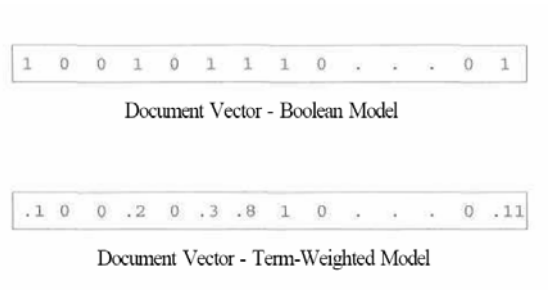

Figure 2. Example of a document vector in binary format (top) and term-weighted format (bottom).

Vector-based representations are sometimes referred to as a 'bag of words' (Pierre, Paolo and Padhraic, 2003), emphasizing that document vectors are invariant with respect to term permutations, since the original word order ω(1), . . . , ω(|v|) is clearly lost. Representations of this kind are appealing for their simplicity. Moreover, although they are necessarily lossy from an information theoretic point of view, many text retrieval and categorization tasks can be performed quite well in practice using the vector-space model. Note that typically the total number of terms in a set of documents is much larger than the number of distinct terms in any single document, $|V| >> |d|$, so that vector-space representations tend to be very sparse. This property can be advantageously exploited for both memory storage and algorithm design.

3.2. Term Weighted Vector Space Model

In Boolean vector models each coordinate of a document vector is zero (when the corresponding attribute is absent) or unity (when the corresponding attribute is present). Term weighting is a widely used refinement of Boolean models that takes into account the frequency of appearance of attributes (such as keywords and keyphrases), the total frequency of appearance of each attribute in the document set, and the location of appearance (e.g., in the title, section header, abstract, or text) (Berry, 2003).

An important family of weighting schemes (Pierre, Paolo and Padhraic, 2003) combines term frequencies (which are relative to each document) with an 'absolute' measure of term importance called inverse document frequency (IDF). IDF decreases as the number of documents in which the term occurs increases in a given collection. So terms that are globally rare receive a higher weight.

Formally, let $D = \{d_1, d_2, ..., d_n\}$ be a collection of documents and for each term ω_j let n_{ij} denote the number of occurrences of ω_j in d_i and n_j the number of documents that contain ω_j at least once. Then we define

$$TF_{ij} = \frac{n_{ij}}{|d_i|},\ IDF_j = \log\frac{n_j}{n}$$

Here the logarithmic function is employed as a damping factor.

The TF-IDF weight of ω_j in d_i can be computed as $x_{ij} = TF_{ij} \cdot IDF_j$

or, alternatively, as

$$x_{ij} = \frac{TF_{ij}}{\max_{\omega_k \in d_i} TF_{ik}} \cdot \frac{IDF_j}{\max_{\omega_k \in d_i} IDF_k}$$

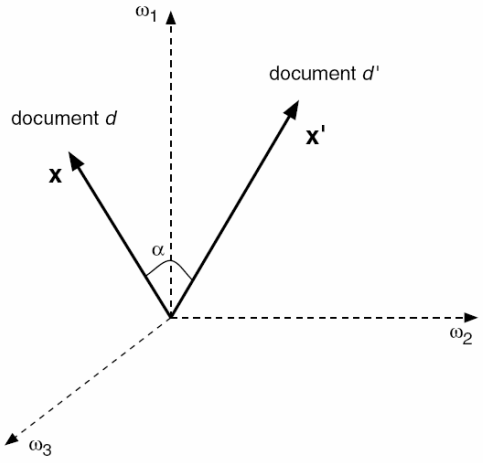

Figure 3. Cosine measure of document similarity.

Document retrieval is now accomplished by computing the similarity between a query vector, q, and a document vector, d, (this measure is simply the cosine of the angle formed by the vector-space representations of the two documents, d and q) using the formula:

$$sim(q,d) = \frac{\sum_t x_{t,d} \cdot x_{t,q}}{\sqrt{\sum_t x_{t,d}^2} \sqrt{x_{t,q}^2}}$$

and then ranking the found documents in decreasing order with respect to this measure (Jackson, and Moulinier, 2002).

4. EVALUATION MEASURES: RECALL AND PRECISION

Two performance metrics gained currency in the 1960s, when researchers began performing comparative studies of different indexing systems. These are recall and precision, and they can be defined as follows.

Let us consider a collection of *n* documents D. Each document is represented by an *m*-dimensional vector, where m = |V | and V is the set of terms that occurred in the collection. Let $q \in R^m$ denote the vector associated with a user query (terms that are present in the query but not in V will be stripped off). Each document is then assigned a score, relative to the query, by computing similarity function $s(x_i, q)$, *i* = 1, …, *n*. The set R of retrieved documents that are presented to the user can be formed by collecting the top-ranking documents according to the similarity measure. The quality of the returned collection can be defined by comparing R to the set of documents R* that is actually relevant to the query.

Two common metrics for comparing R and R* are precision and recall. Precision π is defined as the fraction of retrieved documents that are actually relevant. Recall ρ is defined as the fraction of relevant documents that are retrieved by the system.

More precisely,

$$\pi = \frac{|R \bigcap R^*|}{|R|}, \ \rho = \frac{|R \bigcap R^*|}{|R^*|}$$

Note that in this context the ratio between relevant and irrelevant documents is typically very small. For this reason, other common evaluation measures like accuracy or error rate, where the denominator consists of |D|, would be inadequate (it would suffice to retrieve nothing to get very high accuracy). Sometimes precision and recall are combined into a single number called F_β measure defined as

$$F_\beta = \frac{(\beta^2 + 1)\pi\rho}{\beta^2\pi + \rho}$$

To note that the F1 measure is the harmonic mean of precision and recall. If β tends to zero (∞) the F_β measure tends to precision and if β tends to ∞ the F_β measure tends to recall (Pierre, Paolo and Padhraic, 2003).

5. CONCLUSIONS

This paper has presented the basics of information retrieval: the vector space model for document representation with Boolean and term weighted models, ranking methods based on the cosine factor and evaluation measures: recall, precision and combined measure.

REFERENCES

Berry, M. W. (2003), *Survey of Text Mining: Clustering, Classification, and Retrieval*, Springer, New York, USA.

Chakrabarti, S. (2003), *Mining the Web: Discovering Knowledge from Hypertext Data*, Morgan Kaufmann Publishers, San Francisco, USA.

Frakes, W. B. and R. A. Baeza-Yates (1992), *Information Retrieval: Data Structures & Algorithms*, Prentice-Hall**,** Englewood Cliffs, New Jersey**,** USA.

Jackson, P. and I. Moulinier (2002), *Natural Language Processing for Online Applications Text Retrieval, Extraction and Categorization*, John Benjamins B.V., Amsterdam, The Netherlands.

Manning, C. and H. Schütze (1999), *Foundations of Statistical Natural Language Processing*, MIT Press. Cambridge.

Manning, C. D., P. Raghavan and H. Schütze (2007), *Introduction to Information Retrieval*, Cambridge University Press, Cambridge, England.

Pierre B., F. Paolo and S. Padhraic (2003), *Modeling the Internet and the Web: Probabilistic Methods and Algorithms*, JohnWiley & Sons Ltd, West Sussex, England.