Term Vector Theory and Keyword Weights

Abstract – This is Part 1 of an introductory tutorial series on Term Vector Theory as used in Information Retrieval and Data Mining. The concepts of local and global term weights are briefly presented and the idea of keyword density as a useful weighting scheme for ranking documents is debunked.

Keywords: vectors term vector theory, term weights, local weights, global weights, keyword density, salton
Published: 10-27-2006; Updated: 11-12-2016

Note: This article is part of a legacy series that the author published circa 2006 at http://www.miislita.com, now a search engine site. It is now republished in pdf format here at http://www.minerazzi.com, with its content edited and updated. The original articles can be found referenced in online research publications on IR and elsewhere.

Introduction

Information retrieval (IR) systems frequently assign weights to terms by considering

1. local information from individual documents.
2. global information from a collection of documents.

A traditional weighting scheme is the Vector Space Model pioneered by Salton (Salton, Wong, & Yang, 1975; Salton, 1983; Stata, Bharat, & Maghoul, 1999; Ackermann, 2003; Baeza-Yates & Ribeiro-Neto, 1999; Rijsbergen, 2004; Garcia, 2004; Grossman & Frieder, 2004).

Salton's Vector Space Model

In the vector space model theory, the weight of a term i in a document j is commonly defined as

$$w_{i,j} = f_{i,j} log\left(\frac{D}{d_i}\right) \quad (1)$$

where $f_{i,j}$ is a local term weight and where

- $f_{i,j}$ = frequency or number of times that term *i* occurs in document *j*.
- d_i= document frequency or number of documents that mention term *i.*
- D = number of documents in a database.

Equation (1) is then used to construct a vector of term weights representing document *j*. The query is treated like another document.

Local Weights

In (1), $w_{i,j}$ increases with $f_{i,j}$. This makes the model vulnerable to term repetition abuses, an adversarial IR practice known as keyword spamming or *spamdexing.* So given a query,

- documents of equal lengths and with more instances of *q* are favored during retrieval.
- longer documents mentioning *q* tend to consist of words somehow relevant to *q*.

Global Weights

In (1), the $log(D/d_i)$ term is known as the *inverse document frequency* (IDF_i). In a strict sense, *IDF* is a measure of specificity; i.e., of the discriminatory power of a term. In Parts 2 and 3 of this series, we discuss local and global weights. We also show that (1) is just one of the many vector space models developed by Salton and others (Garcia, 2016a; 2016b).

Generally speaking, the d_i/D ratio is the probability p_i that a document from *D* mentions term *i*. In (1) we have inverted p_i to avoid negative signs.

Equation (1) shows that $w_{i,j}$ decreases as d_i increases. For example, if in a 1000-document database only 10 documents mention "pet", *IDF* = log(1000/10) = 2. However, if only one document mentions this term, *IDF* = log(1000/1) = 3. This means that frequently used terms like "a", "the", and "of", weigh less than rarely used terms.

That makes sense since frequently used terms can hardly be used to discriminate between documents. In general, good query terms are those whose vectors are not too close or distant from document vectors. Let us address this point.

If uncommon query terms are found in the documents, the system will certainly rank these high, but at the expense of returning fewer documents. This can be meaningless. A document ranked, for instance, in position 1 out of 50 documents not necessarily is more relevant than one

ranked in position 10 out of 5,000, 000. This also tells us nothing about the usefulness of uncommon terms. For instance, when querying a commercial search engine like Google or Bing, average Web users do not tend to search for rare terms.

Keyword Density Values

From (1), it is evident that keyword weights are affected by

1. local term counts.
2. the number of documents in a database.

Therefore, the idea that a "keyword density" value can be taken for the weight of a term is misleading. Keyword density is defined as

$$KD_{i,j} = \frac{f_{i,j}}{L_j} \tag{2}$$

where L_j is the length of document j, computed as its total number of terms. So $KD_{i,j}$ is a local weight representing the probability of finding term i in a piece of text of length L. For instance, if a 500-word document mentions “pet” five times, KD = 5/500 = 0.01 or 1%. This result tells us nothing about the position and dispersion of “pet” in this document.

In general, (2) does not prove how specific terms are related to topics or subtopics. We must keep in mind that term distribution can affect text semantics, and even the perception of relevance.

Unfortunately, many search engine optimization/marketing specialists (SEOs/SEMs) waste their time computing (2) with *Keyword Density Tools*, with some going to the extreme of computing localized values in page identifiers and descriptors (e.g., urls, titles, paragraphs, etc). Some of them have claimed in discussion forums and sites that keeping documents within an "optimum" keyword density value affects the way commercial search engines rank web documents.

Keyword Density Failures

Equation (2) tells nothing about the semantic weight of terms in relation to other terms, within a document or collection of documents.

Frankly, SEOs/SEMs that spend their time adjusting keyword density values, going after keyword weight tricks or buying the latest "keyword density tools" are just wasting their time and money.

According to (2), a term equally repeated in two different documents of same length has the same keyword density, regardless of the content of the documents. So if we assume that keyword density values can be taken for keyword weights, then we are

1. ignoring the shear volume of information that a query retrieves.
2. assigning term weights without regard for term relevancy.
3. assigning weights without considering the nature of the queried database.

Points 1 - 3 are contrary to Salton's Vector Space Model. According to (1), term weights are neither word ratios nor they are disconnected from the queried database. Often, a given term equally repeated in two different documents of same length, regardless of content, is weighted differently in the same collection, in different databases, or over time.

Foolish Thinking

If a search marketer wants to compute term weights, he/she may need to replicate the weighting scheme of the target system. But, this is not an easy task since:

- $f_{i,j}$ and IDF_i are defined differently across IR systems.
- if using (1), he/she must know D, total number of documents in the database, and d_i, number of documents containing the queried term.
- number of documents containing the queried term(s) is not the same as the number of documents retrieved, some of which might be relevant or irrelevant.
- To avoid spamdexing, search engines do not publish their algorithms.
- Some search engines may not use Salton's Vector Space Model at all.

Last, but not least, an IR system or commercial search engine may use a variant of Salton's Vector Space Model, combined with other scoring schemes to account for things like link citation and web graph connectivity weights (Broder et al., 1999; Henzinger et al., 1999; Mukherjea,1999; Rafiei & Mendelzon, 1999), and more recently, social weights.

Conclusion

We have presented Part 1 of an introductory tutorial series on Term Vector Theory. The concepts of local and global term weights have been presented.

The idea of keyword density as a useful weighting scheme for ranking documents has been debunked. Keyword density values should not be taken for term weights. Thinking otherwise is foolish (Garcia, 2005).

References

Ackerman, R. (2003). Vector Model Information Retrieval. Retrieved from http://www.hray.com/5264/math.htm

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). Modern Information Retrieval. Adisson Wesley. Book Review. Retrieved from http://www.amazon.com/gp/customer-reviews/R2HC8ULDSMXKZQ/ref=cm_cr_arp_d_rvw_ttl?ie=UTF8&ASIN=020139829X

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. (2000). Graph Structure in the Web. Retrieved from http://www9.org/w9cdrom/160/160.html

Garcia, E. (2004). Term Vector Theory and Keyword Weights. Search Engine Watch forums. Retrieved from http://forums.searchenginewatch.com/showthread.php?t=489

Garcia, E. (2005). The Keyword Density of Non-Sense. E-Marketing News. Retrieved from http://www.e-marketing-news.co.uk/Mar05/garcia.html

Garcia, E. (2016a). The Binary and Term Count Models. Retrieved from http://www.minerazzi.com/tutorials/term-vector-2.pdf

Garcia, E. (2016b). The Classic TF-IDF Vector Space Model. Retrieved from http://www.minerazzi.com/tutorials/term-vector-3.pdf

Grossman, D. A., Frieder, O. (2004). Information Retrieval: Algorithms and Heuristics. Springer. Book Review. Retrieved from http://www.amazon.com/review/RACNGPXD2GNE7/ref=cm_cr_dp_title?ie=UTF8&ASIN=1402030045&channel=detail-glance&nodeID=283155&store=books

Henzinger, M. R., Heydon, A., Mitzenmacher, M., and Najork, M. (2000). On Near-Uniform URL Sampling. Retrieved from http://www9.org/w9cdrom/88/88.html

Mukherjea, S. (2000). WTMS: A System for Collecting and Analyzing Topic-Specific Web Information. Retrieved from http://www9.org/w9cdrom/293/293.html

Rafiei, D. and Mendelzon, A. O. (2000). What is this Page Known for? Computing Web Page Reputations. Retrieved from http://www9.org/w9cdrom/368/368.html

Rijsbergen, K. (2004). The Geometry of Information Retrieval. Cambridge University Press, UK. Book Review. Retrieved from http://www.amazon.com/review/R3FM04FS4ZDHGC/ref=cm_cr_dp_title?ie=UTF8&ASIN=0521838053&channel=detail-glance&nodeID=283155&store=books

Salton, G. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill.

Salton, G.; Wong, A.; Yang, C. S. (1975). A Vector Space Model for Automatic Indexing. Communications of the ACM 18 (11): 613. Retrieved from http://elib.ict.nsc.ru/jspui/bitstream/ICT/1230/1/soltan_10.1.1.107.7453.pdf see also http://www.bibsonomy.org/bibtex/10a4c67f15a4869634d8e5e39ba3e7113

Stata, R., Bharat, K., and Maghoul, F. (2000). The Term Vector Database: fast access to indexing terms for Web pages. Retrieved from http://www9.org/w9cdrom/159/159.html