The Extended Boolean Model

Abstract –This is Part 6 of a tutorial series on Term Vector Theory. The Extended Boolean Model is discussed. By varying the model p-norm parameter, from $p = 1$ to $p = \infty$, we can vary its ranking behavior from that of a vector space-like to that of a strict Boolean-like.

Keywords: boolean model, extended boolean model, p-norm, and queries, or queries
Published: 10-27-2006; Updated: 06-10-2016

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

In Parts 1-5 of this series on vector space models we described several term weight strategies (Garcia, 2016a; 2016b; 2016c; 2016d, 2016e).

This time we want to focus our attention on one of the most versatile models: The Extended Boolean Model developed by Edward Alan Fox. Back in 2006, we contacted Professor Fox and he generously emailed us references of his outstanding research which are still included in the Reference section. A complete list of references is also available online (Fox, 2016). Before discussing the model, an overview of its predecessor, The Standard Boolean Model, is presented.

The Standard Boolean Model

In the Standard Boolean Model for Information Retrieval, the query is in the form of a Boolean expression which might consists of a set of index terms connected by the Boolean operators AND, OR, and NOT.

The documents retrieved for a given query are those that contain index terms in the combination specified by the query. The model is binary in nature where 1 means a document-query term match event and 0 a non-match event. To illustrate, consider the case of two-term queries such as [q_1 OR q_2] and [q_1 AND q_2]. Table 1 shows that three document classes are

retrieved with the queries: those matching both terms, those matching only one of the terms, and those matching neither term.

Table 1. Boolean Model for Information Retrieval

		Query terms		Similarity with query	
	document	q_1	q_2	[q_1 OR q_2]	[q_1 AND q_2]
Class 1	d_1	1	1	1	1
Class 2	d_2	1	0	1	0
	d_3	0	1	1	0
Class 3	d_4	0	0	0	0

The [q_1 OR q_2] query assumes that Class 1 and 2 documents are equally important and have the same document-query similarity. By contrast the [q_1 AND q_2] query assumes that only Class 1 documents are important while Class 2 and 3 documents are useless.

Queries with the Boolean model are easy to implement and, when well formulated, can produce results with high recall and precision. Relevance feedback can be incorporated to the model (Salton, Fox, & Voorhees, 1983a; 1983b; Salton, Fox, Buckley, & Voorhees, 1983). However, a close inspection at Table 1 reveals that the model has the following drawbacks:

- No weights are assigned to query terms.
- Too many or few documents are retrieved so the size of the results is difficult to control.
- Documents are not ranked in any order of presumed importance to the user.
- All document and query terms are assumed to be equally important.
- With OR searches, documents matching at least one or all of the query terms are considered equally important.
- With AND searches, documents not matching one or none of the query terms are considered equally useless.
- With AND and OR queries, the order of the terms in queries and documents does not matter.

Evidently, the model has too many drawbacks and limitations (Fox, 2001).

The Extended Boolean Model

In his 1981 thesis, Wu applied the *p*-norm concept to Information Retrieval. Few years later, in his 1983 thesis, Fox used the *p*-norm to extend the Boolean and Vector Space models, developing what is nowadays known as the Extended Boolean Model (Fox, 1983; Salton, Fox, & Wu, 1983a; 1983b; Wu, 1981; Salton, Buckley, & Fox, 1983). The practical aspects of the model was the subject of Smith's 1990 thesis (Smith, 1990).

In the Extended Boolean Model, the local and global weights of terms present in a document are first normalized by their maximum scores and their products taken; i.e.

$$L_{i,j} = \frac{f_{i,j}}{max\, f_{i,j}} \tag{1}$$

$$G_i = \frac{IDF_i}{max\, IDF_i} \tag{2}$$

$$w_{i,j} = L_{i,j} G_i = \left(\frac{f_{i,j}}{max\, f_{i,j}}\right)\left(\frac{IDF_i}{max IDF_i}\right) \tag{3}$$

Thus, document term weights adopt values between 0 and 1. For a document j mentioning two terms, $w_{1,j}$ and $w_{2,j}$, the term assignment can be described by a two-dimensional term space, as shown in Figure 1.

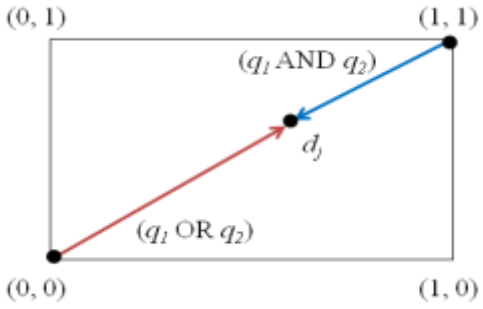

Figure 1. Term space representation of AND and OR two-term queries.

Figure 1 shows that

- for AND queries, the (1, 1) point represents the case where both terms are present in a document.
- for OR queries, (0, 0) point represents the case where both terms are absent from a document.
- the maximum Euclidean distance, $dist_{max}$, between the (0, 0) and (1, 1) points is

$$dist_{max} = \sqrt{(1-0)^2 + (1-0)^2} = \sqrt{2} \tag{4}$$

Therefore, for an AND query, a document that mentions both terms can be represented as a point whose displacement or distance, measured from the (1, 1) point, is

$$dist_{j,AND} = \sqrt{2} - \sqrt{\left(1 - w_{1,j}\right)^2 + \left(1 - w_{2,j}\right)^2} \tag{5}$$

Conversely, for an OR query, a document that mentions one term, but not the other, can be represented in this term space as a point whose distance, $dist_{j,OR}$, measured from the origin of the term space at (0, 0) is less than $dist_{max}$.

$$dist_{j,OR} = \sqrt{\left(w_{1,j} - 0\right)^2 + \left(w_{2,j} - 0\right)^2} = \sqrt{{w_{1,j}}^2 + {w_{2,j}}^2} < \sqrt{2} \tag{6}$$

A document-query similarity measure can then be proposed as a relative distance by normalizing all distances with respect to $dist_{max}$:

$$sim\left(d_j, q_{AND}\right) = \frac{dist_{j,AND}}{dist_{max}} = 1 - \sqrt{\frac{\left(1-w_{1,j}\right)^2 + \left(1-w_{2,j}\right)^2}{2}} \tag{7}$$

$$sim(d_j, q_{OR}) = \frac{dist_{j,OR}}{dist_{max}} = \frac{\sqrt{{w_{1,j}}^2 + {w_{2,j}}^2}}{\sqrt{2}} = \sqrt{\frac{{w_{1,j}}^2 + {w_{2,j}}^2}{2}} \quad (8)$$

We can then rank documents where those close to the (0, 0) point are the least relevant ones.

The *p*-norm

The Euclidean distances measured above are L_2-norms. We can generalize (7) and (8) for *n* numbers of query terms using L*p*-norms, where $1 \leq p \infty$. Hence, (7) and (8) become

$$sim(d_j, q_{AND}) = 1 - \left(\frac{(1-w_{1,j})^p + (1-w_{2,j})^p + \cdots + (1-w_{n,j})^p}{n}\right)^{1/p} \quad (9)$$

$$sim(d_j, q_{OR}) = \left(\frac{{w_{1,j}}^p + {w_{2,j}}^p + \cdots + {w_{n,j}}^p}{n}\right)^{1/p} \quad (10)$$

where the *p* parameter is formally referred to as the *p*-norm.

Combining Boolean Queries

By adopting *p* values between 1 and infinity, we can vary the *p*-norm ranking behavior from a vector-like to a Boolean-like ranking. The *p*-norm then allows for combination of AND/OR queries by recursively grouping operators (Baeza-Yates, Ribeiro-Neto, 1999).

For instance, the query $q = q_1$ AND q_2 OR q_3 can be evaluated recursively by treating the q_1 AND q_2 part as a single term *k* that is part of the OR query $q = k$ OR q_3, and then computing

$$sim(d_j, q_{OR}) = \left(\frac{{w_{k,j}}^p + {w_{3,j}}^p}{2}\right)^{1/p} = \left(\frac{\left(1-\left(\frac{(1-w_{1,j})^p + (1-w_{2,j})^p}{2}\right)^{1/p}\right)^p + {w_{3,j}}^p}{2}\right)^{1/p} \quad (11)$$

Incorporating Query Weights

To reflect the importance of individual terms in a query, term weights are now multiplied by the corresponding query terms. Essentially we are reweighting term weights with query weights. Expressions (9) and (10) then become

$$sim(d_j, q_{AND}) = 1 - \left(\frac{{w_{1,q}}^p (1-w_{1,j})^p + {w_{2,q}}^p (1-w_{2,j})^p + \cdots + {w_{n,q}}^p (1-w_{n,j})^p}{\sum_i^n {w_{i,q}}^p} \right)^{1/p} \quad (12)$$

$$sim(d_j, q_{OR}) = \left(\frac{{w_{1,q}}^p {w_{1,j}}^p + {w_{2,q}}^p {w_{2,j}}^p + \cdots + {w_{n,q}}^p {w_{n,j}}^p}{\sum_i^n {w_{i,q}}^p} \right)^{1/p} \quad (13)$$

When the query terms are fully weighted, ${w_{i,q}}^p = 1$, $\sum_i^n {w_{i,q}}^p = n$, and these similarity measures reduce to the basic expressions (9) and (10).

On a side note, when we reviewed the 2nd Edition of Grossman and Frieder's book (2004) we found a minor typo in page 69 of the text: For the AND query given in (12) the *p*-norm exponent was written inside parentheses, (1 - ${w}^p$).

Implications of setting *p*-norm values

The *p*-norm posses interesting properties that we now discuss. When p = 1

$$sim(d_j, q_{AND}) = sim(d_j, q_{OR}) = \frac{w_{1,q} w_{1,j} + w_{2,q} w_{2,j} + \cdots + w_{n,q} w_{n,j}}{\sum_i^n w_{i,q}} \quad (14)$$

where the distinction between OR and AND queries disappears.

It is clear that (14) is the dot product between document term weights, $w_{n,j}$, and normalized query term weights, $w_{i,q}^* = w_{i,q} / \sum_i^n w_{i,q}$. So by multiplying the coordinates $d_j(w_{1,j}, w_{2,j}, \ldots w_{n,j})$ and $q(w_{1,q}^*, w_{2,q}^*, \ldots w_{n,q}^*)$, and adding together products, documents can be ranked using the similarity coefficient, $simcoeff(d_j, q) = \mathbf{d} \bullet \mathbf{q}$, which is a vector space ranking function.

By contrast, when $p = \infty$, AND queries behave like strict Boolean AND while OR queries behave like strict Boolean OR, where query weights are not weighted. Furthermore, by varying the p-norm, from $p = 1$ to $p = \infty$, we can vary its ranking behavior from that of a vector space-like to that of a strict Boolean-like. Table 2 lists several interpretations, consequences, or justifications for varying p (Fox, 1983; 1996).

Table 2. Interpretation of *p*-values

p-norm	Query	Resulting semantic relation
∞	AND	Strict assignment of query terms as phrases. Document not retrievable unless all query terms are present.
∞	OR	Implements a strict thesaurus. Terms related by OR are substitutable one for another so any query term retrieves the document. Only one of each group of related terms is required for retrieval.
3	AND	Loose query terms as phrases. The presence of all query terms is worth more than the presence of only some of them. Terms are not compulsory.
3	OR	Implements a loose thesaurus. The presence of several terms from a given class is worth more than the presence of only one term.
1	AND, OR	Terms are independent of each other. Distinction between query terms as phrases and thesaurus assignment disappears.

Model Limitations

As noted by Smith (1990), a p-norm model has severe limitations and usability drawbacks. In particular,

- the p-norm model does not satisfy all Boolean algebra properties.
- p-norm retrieval is too slow to be useful.
- formulating p-norm queries is difficult for untrained users.

These are some of the reasons that make the model unattractive for commercial search engines.

Conclusion

In this tutorial, we described The Extended Boolean Model and its *p*-norm as an alternative to traditional Boolean models. By varying the *p*-norm, from $p = 1$ to $p = \infty$, its ranking behavior changes from that of a vector space-like to that of a strict Boolean-like.

When $p = 1$, the difference between AND and OR queries disappears and a vector space model is obtained where a document-query similarity is a similarity coefficient, $simcoeff(d_j, q)$. By contrast, when $p = \infty$ and for unweighted queries, AND queries behave like strict Boolean AND while OR queries like strict Boolean OR.

Although web search engines have incorporated some of the most popular Boolean operators as advanced searches, these are little used by average users. This is unlikely to change, in spite of the many attempts to improve the Extended Boolean Model (Fox & Sharan, 1986; Fox & Koll, 1988; Lee & Fox, 1988; Smith, 1990; Fox, Betrabet, Koushik, & Lee, 1992; Cho, Kim, & Raghavan, 2005; Nguyen, Heo, Lee, Kim, & Whang, 2008; Lv, Zhang, Lou, &Wang, 2015). Search engines can do better by pre-formulating the queries on their own and suggesting these as alternative searches to the end users. Whether this improves performance is still debatable, though.

Exercises

1. Rank the following documents with the Standard and Extended Boolean Models

    - d_1 mentions stock ($w = 0.2$) and market ($w = 0.1$)
    - d_2 mentions stock ($w = 0.5$) and investment ($w = 0.3$)
    - d_3 mentions stock ($w = 0.7$)

    using the following queries where the weight of each term in the query is 1.0.

    - q = stock OR market
    - q = stock OR investment
    - q = market OR investment

2. Repeat previous exercise, replacing OR with AND

References

Baeza-Yates, R. B. Ribeiro-Neto, B. (1999) *Modern Information Retrieval* Chapter 2, p. 40; Addison-Wesley. Book Review. Retrieved from http://www.amazon.com/gp/customer-reviews/R2HC8ULDSMXKZQ/ref=cm_cr_arp_d_rvw_ttl?ie=UTF8&ASIN=020139829X

Cho, J. Kim, M., and Raghavan, V. V. (2005). Adaptive relevance feedback method of extended Boolean model using hierarchical clustering techniques. Retrieved from https://www.researchgate.net/publication/222414600_Adaptive_relevance_feedback_method_of_extended_Boolean_model_using_hierarchical_clustering_techniques

Fox, E. (1983). Extending the Boolean and Vector Space Models of Information Retrieval with P-norm Queries and Multiple Concept Types. Ph.D. Dissertation, Cornell University. Retrieved from https://catalog.hathitrust.org/Record/009232562

Fox, E. (1996). Reviews of the Extended Boolean Information Retrieval Paper. Retrieved from http://ei.cs.vt.edu/~cs5604/f96/art-summs/SALT83.txt

Fox, E. (2001). CS5604 - Information Storage and Retrieval (F2001). Retrieved from http://ei.cs.vt.edu/~cs5604/f96/cs5604cnIF/IF3.html

Fox, E. (2016). Edward A. Fox – CV. Retrieved from http://fox.cs.vt.edu/cv.htm

Fox, E., Betrabet, S., Koushik, M., and Lee, W. (1992). Extended Boolean Models. In Information Retrieval: Data Structures & Algorithms, eds. W. Frakes & R. Baeza-Yates, Prentice-Hall, 1992, 393-418. Retrieved from http://orion.lcg.ufrj.br/Dr.Dobbs/books/book5/chap15.htm

Fox, E. and Koll, M. (1988). Practical Enhanced Boolean Retrieval: Experiences with the SMART and SIRE Systems. IP&M, 24(3): 257-267. Retrieved from https://www.researchgate.net/publication/220229023_Practical_enhanced_Boolean_retrieval_Experiences_with_the_SMART_and_SIRE_systems

Fox, E. and Sharan, S. (1986). A Comparison of Two Methods for Soft Boolean Operator Interpretation in Information Retrieval, TR-86-1, VPI&SU Computer Science Dept., Jan. 1986, Blacksburg, VA. Retrieved from http://eprints.cs.vt.edu/archive/00000008/01/TR-86-01.pdf

Garcia, E. (2016a). Term Vector Theory and Keyword Weights. Retrieved from http://www.minerazzi.com/tutorials/term-vector-1.pdf

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

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

Garcia, E. (2016d). An Introduction to Local Weight Models. Retrieved from http://www.minerazzi.com/tutorials/term-vector-4.pdf

Garcia, E. (2016e). Introduction to Global Weights. Retrieved from http://www.minerazzi.com/tutorials/term-vector-5.pdf

Grossman, D. A., Frieder, O. (2004). *Information Retrieval: Algorithms and Heuristics.* Chapter 2, p. 69. 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

Lee, W. and Fox, E. (1988). Experimental Comparison of Schemes for Interpreting Boolean Queries, TR-88-27, VPI&SU Comp. Science Dept. Blacksburg, VA. Retrieved from http://eprints.cs.vt.edu/archive/00000112/01/TR-88-27.pdf

Lv, F., Zhang, J., Lou, J., and Wang, S. (2015). CodeHow: Effective Code Search Based on API Understanding and Extended Boolean Model (E). Automated Software Engineering (ASE), 2015 30th IEEE/ACM International Conference. Retrieved from http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=7372014

Nguyen, T., Heo, J., Lee, Kim, Y., and Whang, K. (2008). Query Expansion Using Augmented Terms in an Exended Boolean Model. Journal of Computing Science and Engineering, Vol. 2, No. 1, 26-43. Retrieved from http://jcse.kiise.org/posting/2-1/jcse_2-1_13.pdf

Salton, G., Buckley, C., and Fox, E. (1983). Automatic Query Formulations in Information Retrieval. JASIS, 1983, 34(4): 262-280. https://ecommons.cornell.edu/bitstream/handle/1813/6363/82-524.pdf?sequence=1&isAllowed=y

Salton, G., Fox, E., and Wu, H. (1983a). An Automatic Environment for Boolean Information Retrieval. In Information Processing 83 (Proc. 1983 IFIP Paris Congress), R.E.A. Mason (ed.), North-Holland, 1983, 755-762. Retrieved from https://www.researchgate.net/publication/221330637_An_Automatic_Environment_for_Boolean_Information_Retrival

Salton, G., Fox, E., and Wu, H. (1983b). Extended Boolean Information Retrieval. Communications of the ACM, 1983, 26(12): 1022-1036. Retrieved from http://neuron.csie.ntust.edu.tw/homework/93/Fuzzy/%E6%97%A5%E9%96%93%E9%83%A8/homework_1/D9009204/Fuzzy%20Homework%20I.files/p1022-salton.pdf
See also https://ecommons.cornell.edu/bitstream/handle/1813/6351/82-511.ps?sequence=2

Salton, G., Fox, E., and Voorhees, E. (1983a). Advanced Feedback Methods in Information Retrieval. TR 83-570. Cornell Univ. Dept. of Computer Science, NY. Retrieved from https://ecommons.cornell.edu/bitstream/handle/1813/6410/83-570.ps?sequence=2

Salton, G., Fox, E., and Voorhees, E. (1983b). A Comparison of Two Methods for Boolean Query Relevance Feedback. TR 83-564, Cornell Univ. Dept. of Computer Science, NY. Retrieved from http://ecommons.cornell.edu/bitstream/handle/1813/6404/83-564.ps?sequence=2

Salton, G., Fox, E. A., Buckley, C., and Voorhees, E. (1983). Boolean Query Formulation with Relevance Feedback. TR 83-539, Cornell Univ. Dept. of Computer Science, NY. Retrieved from https://ecommons.cornell.edu/bitstream/handle/1813/6379/83-539.pdf?sequence=1

Smith, M. E. (1990). Aspects of the *p*-Norm Model of Information Retrieval: Syntactic Query Generation, Efficiency, and Theoretical Properties. Ph.D. Dissertation, Cornell University. Retrieved from https://ecommons.cornell.edu/handle/1813/6968

Wu, H. C. C. (1981). On Query Formulation in Information Retrieval. Ph.D. Dissertation, Cornell University. Retrieved from http://dl.acm.org/citation.cfm?id=909955