Расширенная булева модель поиска

Р. В. Шарапов, Е. В. Шарапова, Т. Е. Меркулова
Муромский институт Владимирского государственного университета, Муром, Россия

Булева модель стала использоваться в информационно-поисковых системах достаточно давно. Это одна из старейших моделей поиска. Основным достоинством такой модели является ее простота, способность работать с большими объемами информации и высокая скорость выполнения поисковых запросов. По этой причине на основе булевой модели построено большое количество поисковых систем.

В булевой модели запросы пользователей представляют собой логические выражения, в которых слова связаны операторами AND, OR и NOT. Для того чтобы документ был найден, в нем должны содержаться все слова, связанные оператором AND, или хотя бы одно из слов, связанных оператором OR. Не трудно заметить, что при сложных запросах, состоящих из нескольких слов, и большом количестве документов в поисковой базе может наблюдаться некий дисбаланс результатов поиска:

- список результатов поиска при использовании оператора AND может оказаться слишком коротким, так как из результатов поиска исключаются все документы, в которых отсутствует хотя бы одно из слов запроса;
- список результатов поиска при использовании оператора OR может оказаться слишком большим, так как в результаты поиска включаются все документы, в которых встречается хотя бы одно из слов запроса.

Кроме того, у булевой модели есть существенный недостаток – в ней нет возможности установить веса термов (слов) и, соответственно, нельзя провести ранжирование результатов поиска. По сути, при поиске документы делятся на две группы – соответствующие и несоответствующие запросу. Так, при использовании оператора AND документы, не содержащие по крайней мере одного из слов запроса, являются столь же несоответствующими запросу, как и документы, не содержащие ни одного из слов запроса. Аналогично при

использовании оператора OR: документы, содержащие одно из слов запроса, в равной степени соответствуют запросу, как и документы, содержащие все слова запроса.

Из-за этого современные информационно-поисковые системы практически перестали строиться на основе булевой модели. Современные системы чаще всего используют различные варианты векторной модели, которые позволяют производить ранжирование результатов поиска, обладают неплохими скоростными характеристиками, но требуют большего числа вычислений. Многие информационно-поисковые системы, использующие булеву модель, в настоящее время утрачивают конкурентоспособность. Перевод же на другие модели поиска означает практически полную их замену. Процедура эта довольно трудоемкая и дорогостоящая. По этой причине интерес представляет модификация существующей булевой модели для обеспечения дополнительной гибкости систем.

Решением проблемы является расширение булевой модели, дающее возможность назначать весами термов, осуществлять поиск с частичным соответствием и производить ранжирование результатов поиска.

Впервые расширенная модель была предложена Дж. Солтоном и Е. Фоксом [Salton, and et. 1983; Baeza-Yates, and et. 1999]. Основная идея состоит в комбинации булевой формулировки запроса и элементов векторной модели. В сущности, предлагается расширить булеву модель элементами векторной модели. Эта модифицированная модель получила название расширенной булевой модели (Extended Boolean Model) [Baeza-Yates, and et. 1999]. В дальнейшем модель была дополнена другими исследователями [Wong, and et. 1988; Paice 1984, Fox, and et. 1986].

Рассмотрим конъюнктивный (оператор AND, пересечение) логический запрос $q = k_x \wedge k_y$. Согласно булевой модели, если документ содержит только терм k_x или терм $k_{y,}$ то он не соответствует запросу так же, как документ, не содержащий ни одного из термов. Однако такой бинарный критерий выбора из

двух альтернатив является недостаточно гибким. Аналогичная ситуация наблюдается при дизъюнктивных запросах (оператор OR, объединение).

Если рассматриваются только два слова в запросе (терма), то можно отобразить запросы и документы на двумерной карте (рисунок 1) [Baeza-Yates, and et. 1999]. Документ d_j позиционируется в этом пространстве на основе весов $w_{x,j}$ и $w_{y,j}$, связанных с парами $[k_x, d_j]$ и $[k_y, d_j]$. После проведения нормализации, эти веса будут принимать значения между 0 и 1. Например, эти веса могут вычисляться как TF-IDF коэффициенты (аналогично векторной модели):

$$w_{x,j}=tf_{norm\,x,j} \times idf_{norm\,x},$$

$$tf_{norm\,i,j} = \frac{tf_{i,j}}{tf_{\max x,j}},$$

$$idf_{norm\,x} = \frac{idf_x}{idf_{\max x}},$$

где $tf_{norm\,x,j}$ - нормализованная частота терма k_x в документе d_j;

$idf_{norm\,x}$ - нормализованная инверсная частота документов для терма k_x;

$tf_{x,j}$ - частота терма k_x в документе d_j;

$tf_{max\,x,j}$ - максимальная частота терма k_x в документе d_j;

idf_x - инверсная частота документов для терма k_x;

$idf_{max\,x}$ – максимальная инверсная частота документов для терма k_x;

$f_{x,j}$ - нормализованная частота терма k_x в документе d_j.

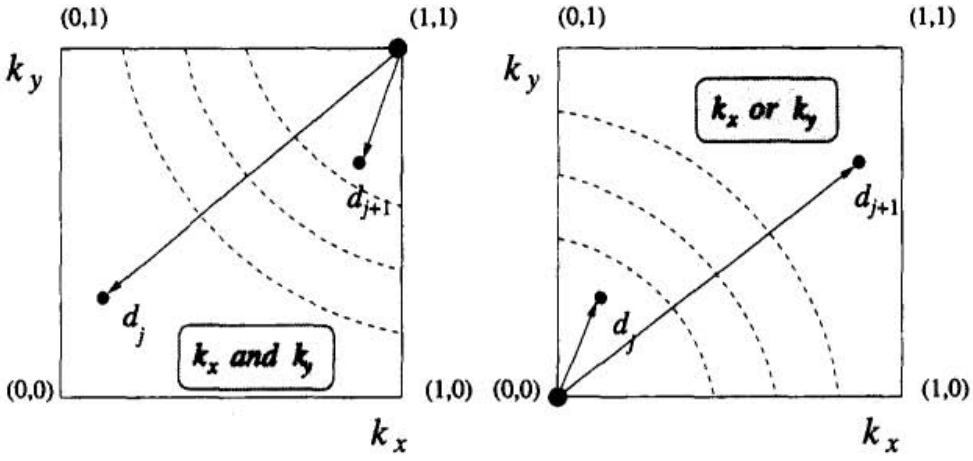

Рис. 1 Пространство, составленное из двух термов k_x и k_y

На рисунке видны две интересные особенности. Во-первых, для дизъюнктивного запроса $q_{or} = k_x \vee k_y$ точка (0,0) является самой нежелательной. Это позволяет считать расстояние от (0,0) как меру подобия запросу q_{or}. Во-вторых, для конъюнктивного запроса $q_{and} = k_x \wedge k_y$ точка (1,1) является самой желательной. Это дает возможность считать расстояние от точки (1,1) как меру подобия запросу q_{and}. Кроме того, такие расстояния могут быть нормализованы. В результате меры подобия документа запросу будут выглядеть следующим образом [Baeza-Yates, and et. 1999]:

$$sim(q_{or}, d) = \sqrt{\frac{w_{x,j}^2 + w_{y,j}^2}{2}}$$

$$sim(q_{and}, d) = 1 - \sqrt{\frac{(1 - w_{x,j})^2 + (1 - w_{y,j})^2}{2}}$$

Если веса – логические (то есть $w_{x,j} \in \{0,1\}$), то документ будет располагаться в одном из четырех углов – (0,0), (0,1), (1,0) или (1,1). Тогда значения для $sim(q_{or},d)$ ограничены 0, $1/\sqrt{2}$ и 1. Аналогично, значения для $sim(q_{and},d)$ ограничены 0, $1-1/\sqrt{2}$ и 1. Если учесть, что число термов в коллекции документов равно *t,* булева модель может быть расширена, чтобы рассматривать евклидовы расстояния в t-мерном пространстве.

Проводя обобщение, можно принять теорию векторных норм. Модель *p*-нормы обобщает понятие расстояния, включающего не только евклидовы расстояния, но и *p*-расстояния, где $1 \le p \le \infty$ – вновь введенный параметр, значение которого определяется во время формирования запроса. Обобщенный дизъюнктивный запрос тогда можно представить как

$$q_{or} = k_1 \vee^p k_2 \vee^p \ldots \vee^p k_m$$

Аналогично, обобщенный конъюнктивный запрос можно представить как

$$q_{and} = k_1 \wedge^p k_2 \wedge^p \ldots \wedge^p k_m$$

Тогда функции соответствия документов запросу будут

$$sim(q_{or}, d_j) = \left(\frac{w_{1,j}^p + w_{2,j}^p + ... + w_{m,j}^p}{m} \right)^{1/p}$$

$$sim(q_{and}, d_j) = \left(\frac{(1 - w_{1,j})^p + (1 - w_{2,j})^p + ... + (1 - w_{m,j})^p}{m} \right)^{1/p}$$

P-норма обладает несколькими интересными свойствами [Baeza-Yates, and et. 1999]. Во-первых, когда *p=1*, получаем

$$sim(q_{or}, d_j) = sim(q_{and}, d_j) = \frac{w_{1,j} + w_{2,j} + ... + w_{m,j}}{m}$$

Во-вторых, при *p*=∞ можно записать

$$sim(q_{or}, d_j) = max(w_{i,j})$$

$$sim(q_{and}, d_j) = min(w_{i,j})$$

Таким образом, для *p=1* конъюнктивные и дизъюнктивные запросы определяются суммой весов термов в документах, подсчитанных на основе формул подобия в векторном пространстве. Для *p*=∞ запросы оцениваются согласно терминам нечеткой логики. Изменяя параметр *p* между 1 и бесконечность, мы можем изменить *p*-норму, варьируя ранжирование между векторной и булевой моделями. Это весьма хороший аргумент в пользу использования расширенной булевой модели.

Обработка более общих запросов выполняется путем группировки операторов в предопределенном порядке. Например, рассмотрим запрос $q = (k_1 \wedge^p k_2) \vee^p k_3$. Подобие *sim(q, d_j)* между документом d_j и этим запросом вычисляется как

$$sim(q, d) = \left(\frac{\left(1 - \left(\frac{(1 - w_{1,p})^p + (1 - w_{2,p})^p}{2} \right)^{1/p} \right)^p + w_{3,j}^p}{2} \right)^{1/p}$$

Эта процедура может быть применена рекурсивно независимо от числа операторов AND/OR. Еще одна интересная особенность расширенной булевой модели – возможность использования комбинаций различных значений параметра *p* в том же самом запросе. Например, запрос

$$(k_1 \vee^2 k_2) \wedge^\infty k_3$$

показывает, что k_1 и k_2 должны считаться как в векторной модели, но k_3 должна присутствовать обязательно (то есть конъюнкция интерпретируется как логическая операция).

Надо заметить, что расширенная булева модель ослабляет булеву алгебру, интерпретирующую логические операции в терминах алгебраических расстояний. В этом смысле это – гибридная модель, которая включает свойства и теоретических, и алгебраических моделей.

Приведенные модификации дают возможность обойти ограничения булевой модели и дать новые возможности ее использования в информационно-поисковых системах. Появившаяся возможность осуществления ранжирования делает расширенную булеву модель более конкурентоспособной.

Список литературы

Salton, and et. 1983 – *Salton, G.* Extended Boolean Information Retrieval. Communications of the ACM / G. Salton, E. Fox, H. Wu. – 1983. – 26(11). – P. 1022-1036.

Baeza-Yates, and et. 1999 – *Baeza-Yates, R.* Modern information retrieval / R. Baeza-Yates, B. Ribeiro-Neto. – Addison Wesley : ACM Press Books, 1999

Frakes, and et. 1992 – *Frakes, W.* Information Retrieval: Data Structures & Algorithms / W. Frakes, R. Baeza-Yates. – Prentice-Hall, 1992.

Wong, and et. 1988 – *Wong, S. K. M*. Extended Boolean Query Processing in the Generalized Vector Space Model / S. K. M. Wong, W. Ziarko, V. V. Raghavan, P. C. N. Wong // Information Systems. – 1988. – 14(1). – P. 47-63.

Paice 1984 – *Paice, C. P.* Soft Evaluation of Boolean Search Queries in Information Retrieval Systems / C. P. Paice // Information Technology, Res. Dev. Applications. – 1984. – 3(1). – P. 33-42.

Fox, and et. 1986 – *Fox, E. A.* Comparison of Two Methods for Soft Boolean Interpretation in Information Retrieval. Technical Report TR-86-1 / E. A. Fox, S. A. Sharat. – Virginia Tech, Department of Computer Science, 1986.

An expanded Boolean model for searching

Ruslan V. Sharapov, Ekaterina V. Sharapova, Tamara E. Merkulova

Murom Institute of Vladimir State University, Murom, Russia

This paper considers opportunities for updating Boolean models for retrieval of search results. Information about expanded Boolean models and opportunities for application in information retrieval systems are given.