УДК 681.3.016(075)

ОЦЕНКА КАЧЕСТВА ПОИСКОВЫХ СИСТЕМ

д.т.н., проф. Е.И. Бобыр, И.Е. Лещенко

Математические модели позволяют значительно упростить алгоритмы и программы, реализующие поиск, оценить качество поиска и сравнить по заданным критериям различные поисковые системы. Для описания моделей поиска на основе семантических сетей предлагается использовать метод производящих функций.

Введение. Сетевые модели представления знаний различаются между собой типами используемых отношений (связей) и делятся на классифицирующие, функциональные и семантические сети [1].

Известно, что приблизительно 3 из 5 поисковых систем разработаны эмпирически, без предварительного абстрактного моделирования процесса поиска на базе каких-либо математических моделей [2]. Вместе с тем математические модели процесса поиска желательны даже в простейших системах, поскольку они позволяют значительно упростить алгоритмы и программы, реализующие поиск.

Кроме того, как только речь заходит о повышении качества и надежности поиска, о большом объеме отыскиваемой информации, о потоке пользовательских запросов, кроме эмпирически разработанных алгоритмов поиска применение математических моделей процессов поиска является полезным, а в целом ряде случаев и необходимым. В настоящее время математические модели процессов поиска информации применительно к сложным семантическим сетям далеки от совершенства и требуют дальнейшей проработки.

Традиционные подходы к организации поиска информации можно разделить на три группы: методы индексного поиска, статистические методы и методы, основанные на базах знаний. Соответственно и все многообразие моделей информационного поиска принято делить на три вида: теоретико-множественные (булевская, нечетких множеств, расширенная булевская), алгебраические (векторная, обобщенная векторная, латентно-семантическая, нейросетевая) и вероятностные [2].

Целью данной статьи является поиск метода, позволяющего достаточно просто получить основные характеристики поисковых систем на базе семантической сети для поиска информации в базах данных.

Модель поиска – это некоторое упрощение реальности, на основании которого получаются зависимости, позволяющие алгоритму принять решение: какой документ считать найденным с заданным уровнем надежности (вероятности) и как его необходимо ранжировать.

Однако какова бы не была модель, поисковая система нуждается в оценке качества поиска. Ибо именно благодаря оценке качества можно говорить о применимости или неприменимости той или иной модели для конкретной проблемной области и сравнивать различные варианты построения поисковых систем. Основные критерии оценки качества построения поисковых систем – время поиска (скорость), точность (вероятность правильного поиска) и полнота (релевантность) ответов.

Основой формализации семантических знаний о предметной области часто является направленный граф с помеченными вершинами и дугами – семантические сети [1]. Вершинам ставятся в соответствие конкретные объекты, а дугам – семантические отношения между ними. Метки вершин имеют ссылочный характер и представляют собой некоторые имена. В роли имен могут выступать, например, слова естественного языка. Метки дуг обозначают элементы множества отношений.

Для исследования зависимости времени поиска и вероятности правильного поиска от заданных пороговых значений необходимо, прежде всего, выбрать математический аппарат, позволяющий достаточно просто анализировать сложные процессы. Очень удобным математическим аппаратом для математических моделей на основе семантических сетей представляемых в виде направленных графов, является метод производящих функций [3].

При использовании этого метода составляется вероятностно-временной граф (ВВГ), описывающий функционирование системы поиска. Пары (P_{ij}, t_{ij}) определяют вероятность выбора дуги ij (P_{ij}) и время ее прохождения (t_{ij}). Вводится функция дуги $f(P_{ij}, t_{ij})$. Вид этой функции должен быть таким, чтобы при нахождении произведений функции вероятности P_{ij} перемножались, а время суммировалось. Этим условиям удовлетворяет функция

$$f_{ij}(P_{ij}, \quad t_{ij}) = P_{ij} z^{t_{ij}}, \tag{1}$$

где z – параметр. Тогда функция последовательности k_g дуг может быть записана в виде

$$f_{1,2,\ldots,k_g}(z) = \prod_{i=1}^{k_g} P_{i,i+1} z^{t_{i,i+1}}. \tag{2}$$

Производящая функция $F(z)$, соответствующая графу, есть сумма

функции всех путей, соединяющих начальную и конечную вершины графа. Так как конечная вершина графа может быть разделена на две, соответствующие правильному поиску и поиску с ошибкой, то производящая функция записывается в виде: $F(z) = F_{пр}(z) + F_{ош}(z)$, где $F_{пр}(z)$ и $F_{ош}(z)$ – функция дуг, соединяющих начальную вершину и вершины, обозначающие соответственно правильный поиск и поиск с ошибкой. Для упрощения нахождения производящей функции необходимо проводить эквивалентные преобразования исходного графа [3]. Эквивалентные преобразования осуществляются до тех пор, пока можно будет написать функцию, характеризующую переход по графу из начального состояния в конечное, т. е. производящую функцию F(z). Среднее время выполнения процесса поиска, дисперсия и вероятность ошибки при этом определяются из формул [3]:

$$T_{ср} = \left.\frac{dF(z)}{dz}\right|_{z=1};$$

$$D_{T_{ср}} = \left.\frac{d^2F(z)}{dz^2}\right|_{z=1} + \left.\frac{dF(z)}{dz}\right|_{z=1} - \left(\left.\frac{dF(z)}{dz}\right|_{z=1}\right)^2; \qquad (3)$$

$$P_{ош} = F_{ош}(z)\big|_{z=1}.$$

Пусть семантическая сеть представляется графом (рис. 1), где вершины 5, 6 графа соответствуют безошибочному поиску заданной информации, а вершины 4, 7 – поиску с ошибкой.

Согласно [3] при эквивалентных преобразованиях исходного графа последовательно соединенные дуги заменяют одной с функцией, равной произведению функций этих дуг (см. 2), а при параллельном соединении – функцией, равной сумме функций этих дуг, т.е. в общем случае при параллельном соединении дуг

$$f_{1,2,\ldots,kg}(z) = \sum_{i=1}^{kg} P_i z^{T_i}. \qquad (4)$$

Так, параллельное соединение дуг вершин 1 – 2, 1 – 3 на графе рис. 1, эквивалентно дугам с функциями:

$$f_{12}(z) = \sum_{i=1}^{3} P_{12i} Z^{T_{12i}}; \qquad (5)$$

$$f_{13}(z) = \sum_{i=1}^{3} P_{13i} Z^{T_{13i}}. \qquad (6)$$

При эквивалентных преобразованиях последовательно соединенные дуги заменяют одной с функцией, равной произведению функций этих дуг.

Тогда дуги, эквивалентные вершинам 1 – 4, 1 – 5, 1 – 6, 1 – 7, будут представлять собой последовательные соединения, и равняться соответственно:

$$f_{14} = P_{24} Z^{T_{24}} \sum_{i=1}^{3} P_{12i} Z^{T_{12i}}; \quad (7)$$

$$f_{15} = P_{25} Z^{T_{25}} \sum_{i=1}^{3} P_{12i} Z^{T_{12i}}; \quad (8)$$

$$f_{16} = P_{36} Z^{T_{36}} \sum_{i=1}^{3} P_{13i} Z^{T_{13i}}; \quad (9)$$

$$f_{17} = P_{37} Z^{T_{37}} \sum_{i=1}^{3} P_{13i} Z^{T_{13i}}. \quad (10)$$

После проделанных эквивалентных преобразований параллельных и последовательных соединений дуг, граф будет иметь вид рис. 2.

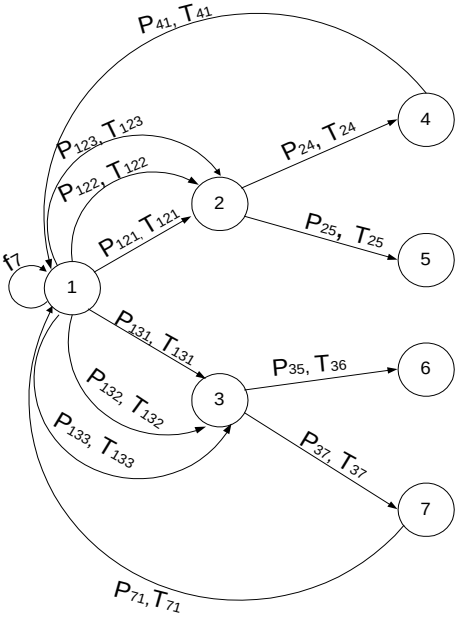

Рис. 1. Исходный граф сети

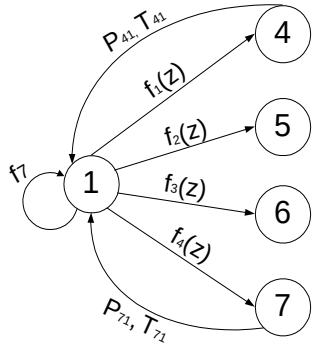

Рис. 2. Граф после промежуточных преобразований

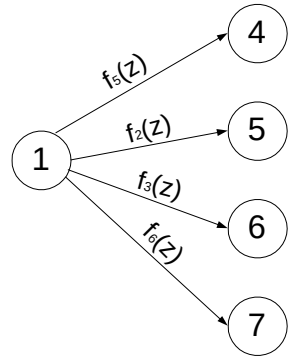

Рис. 3. Граф после эквивалентных преобразований

Если же в графе, как в нашем случае, имеется петля, то это характеризует бесконечно повторяющийся процесс, математически описываемый геометрической прогрессией. Следовательно, при эквивалентных преобразованиях дуга с петлей заменяется одной дугой с функцией

$$f(z) = f_1(z)/[1 - f_2(z)]. \qquad (11)$$

Применительно к рис. 2 знаменатель прогрессии равен функции $f_7(z)$, т. е. функции петли. Тогда последовательное соединение на графе рис. 2 функции петли и дуг 4 – 1 и 7 – 1 можно записать в виде:

$$f^{14}(z) = f_5(z) = \frac{f_{14}(z)}{1 - f_7(z)} P_{41} Z^{T_{41}}; \qquad (12)$$

$$f^{17}(z) = f_6(z) = \frac{f_{17}(z)}{1 - f_7(z)} P_{71} Z^{T_{71}}. \qquad (13)$$

В результате таких преобразований получили граф, эквивалентный исходному (рис. 3), но значительно проще. Тогда производящие функции Z1 (вершины 1 – 4 – 5) и Z2 (вершины 1 – 6 – 7), соответствующие графу, будут иметь вид:

$$Z1 = Z1_{пр} + Z1_{ош} = f_{15} + f^{14}, \quad Z2 = Z2_{пр} + Z2_{ош} = f_{16} + f^{17}.$$

Полученные выражения позволяют применить к Z-функции операцию дифференцирования для получения указанных выше характеристик семантической сети. Среднее время выполнения процесса поиска, дисперсии и вероятности ошибки при этом определяются из соответствующих формул дифференцированием производящих функций Z1 и Z2 эквивалентного графа семантической сети.

Вывод. Использование метода производящих функций для описания математических моделей поиска информации на основе семантических сетей позволяет достаточно просто получить основные характеристики поисковых систем и провести их сравнительный анализ.

ЛИТЕРАТУРА

1. *Искусственный интеллект / В.Н. Бондарев, Ф.Г. Аде. – Севастополь: Изд-во СевНТУ, 2002. – 615 с.*
2. *Как работают поисковые системы / И. Сигалович – Мир Internet, № 10. – 2002.– 9 с. – www.iworld.ru .*
3. *Адаптивная компенсация помех в каналах связи / Ю.И. Лосев и др. – М.: Радио и связь, 1988. – 208 с.*

Поступила 18.08.2003

***БОБЫР Евгений Иванович**, доктор технических наук, профессор, зав. кафедрой информационных технологий и документоведения Харьковского гуманитарного университета «Народная украинская академия». В 1974 году окончил Харьковскую военную инженерную радиотехническую академию ПВО. Область научных интересов – вычислительные системы и сети АСУ и их эффективность и помехозащищенность.*

***ЛЕЩЕНКО Ирина Евгеньевна**, соискатель кафедры информационных технологий и документоведения Харьковского гуманитарного университета «Народная украинская академия». В 1990 году окончила Харьковский институт радиоэлектроники. Область*

научных интересов – программное обеспечение АСУ.