ІНФОРМОЛОГІЯ

УДК: 025.4

ОЦЕНКА ЭФФЕКТИВНОСТИ ПОИСКОВОЙ СИСТЕМЫ ВИРТУАЛЬНОГО МУЗЕЯ НОБЕЛИСТИКИ

И.В.Тявкин

В статье рассмотрено описание методов расчета релевантности для поисковых систем, размещенных в сети Интернет. Приведена математическая модель поиска, включающая ма-тематическое описание, состоящее из критериев оценки степени соответствия дескрипторов и соотношений для вычисления весовых коэффициентов дескрипторов и алгоритм дескрипторного поиска в информационно-поисковой системе виртуального музея нобелистики.

Ключевые слова: информационно-поисковая система, поиск, алгоритм, дескриптор.

The article deals with description of methods of calculation of relevance for the searching systems placed in the Internet. The mathematical model of search including the mathematical description consisting of criteria of estimate of degree of accordance of descriptors and correlations for calculation of weight factors of descriptors and algorithm of descriptor search in an information retrieval system of a virtual museum

Для описания качества работы поисковой системы используются количественные оценки, критерии смыслового соответствия запроса и ресурсов. Это соответствие для несетевых массивов определяется понятием релевантность. На основе оценки релевантности ресурсов запросу оценивается эффективность произведенного информационной системой поиска.

Функционирование поисковой системы характеризуют, в первую очередь, двумя количественными оценками эффективности поиска, а именно полнотой (C) и точностью (P) [1;4;6]. Оценивают полноту и точность по следующим зависимостям:

$$C = \frac{a}{a+c},$$

$$P = \frac{a}{a+b},$$

где a – количество выданных релевантных данных (документов), b – количество выданных нерелевантных данных (документов), c – количество не выданных релевантных данных (документов), оставшихся в БД.

В идеальной информационно-поисковой системе C=P=1. В реальных информационно-поисковых системах коэффициент полноты поиска может достигать значений 0,7-0,9, а коэффициент точности обычно находится в пределах 0,1-1,0 [3].

В большинстве случаев, пользователь не может знать заранее, сколько документов, релевантных введенному его запросу, находится в БД. Поэтому ни пользователь, ни алгоритм информационно-поисковой системы (ИПС) быстро не могут определить коэффициенты полноты и точности.

Для определения эффективности работы алгоритма ИПС и более точной оценки реле-вантности прибегают к оценке эксперта [2]. Но и в оценках релевантности экспертов могут встретиться ошибки. Более того, оценки, выполненные разными экспертами, не всегда совпадают. Релевантность, определяемая на основе использования оценок эксперта, называется экспертной релевантностью. При обработке документов и данных в поисковых системах к экспертной оценке прибегают редко.

Практически во всех алгоритмах работы ИПС заложена оценка соответствия содержания запроса и некоторых формальных характеристик ресурса. Релевантность, находимая таким образом, называется формальной релевантностью [5].

В разработанной нами ИПС "Виртуальный музей нобелистики" (ИПС ВМН) использована формальная релевантность. При проектировании ИПС ВМН стоит задача разработки и реализации алгоритмов поиска по нескольким БД и нахождения наиболее релевантных данных.

ИПС ВМН представляет собой программный продукт, включающий технологии баз данных, информационного поиска, программирования трехмерного пространства. Для пользователя, в проекте ИПС ВМН, реализовано два направления работы. Пользователь выбирает либо виртуальный музей, где может ознакомиться с виртуальными турами ВМН, либо использует поисковую систему для обращения к встроенным БД и последующего поиска интересующей информации.

В ИПС ВМН используются такие же технологии, которые используются в любых информационно-поисковых системах сети Интернет. Другими словами, в окне программы отображено текстовое поле, куда пользователь должен ввести запрос, представляющий собой совокупность дескрипторов (ключевых слов), по которым осуществляется поиск. Большинство ИПС разделяет запрос на дескрипторы путем нахождения пробелов, которые и служат границами дескрипторов.

ИПС ВМН в соответствие с алгоритмом осуществляет поочередный поиск по ключе-вым словам. Поиск выполняется в базах данных, таблицах и столбцах, отмеченных пользователем. Следует отметить, что ИПС осуществит присоединение данных из других таблиц, отмеченных пользователем и выбранных по ключевым полям.

Алгоритм поиска включает в себя следующие этапы:

1) разделение на дескрипторы и анализ поискового запроса;

2) просмотр БД и таблиц, выявление областей (таблиц) БД для последующего поиска;

3) поиск в выбранных столбцах таблиц по дескрипторам запроса;

4) запоминание найденных данных во внутренней таблице ИПС ВМН;

5) поиск ключевых полей в таблице ИПС ВМН;

6) поиск в таблицах БД по выбранным данным из ключевых полей таблицы ИПС ВМН, запоминание данных в таблице ИПС ВМН.

Когда поиск данных завершен, формируется таблица со всеми найденными в БД данными. Далее ИПС ВМН осуществляет формирование отчета. Алгоритм построен так, чтобы все найденные данные проанализировать по критериям (перечислены ниже), выстроить найденные в БД записи по уменьшению степени соответствия данных запросу пользователя.

При анализе данных на соответствие дескрипторам, вычисляются следующие критерии:

1) количество совпавших дескрипторов запроса (Rel – релевантность);

2) количество копий дескрипторов, найденных в документе (если в полях таблиц присутствуют полнотекстовые документы) (Cij – полнота);

3) последовательность дескрипторов, найденных в таблицах (Di – вес найденной информации).

Тогда Q (степень соответствия дескрипторов) вычисляется по формуле:

$$Q = \sum_{i=1}^{n} \left(Rel_i + \sum_{j=0}^{m} C_{ij} + D_i \right)$$

де: Reli – степень соответствия дескриптора (поискового запроса пользователя) выбранным документам или данным в БД, i = 1 … n – номер дескриптора; Cij – полнота, т.е. количество копий дескриптора, где j = 0 … m – количество документов в БД; Di – вес i-го дескриптора поискового запроса (в диапазоне от 0 до 1).

Вес дескрипторов поискового запроса в алгоритме определяется по формуле:

$$L_i = \frac{L_{i-1}}{2}; i = 2...n$$

Вес первого дескриптора является самым большим и равен 1, т.к. обычно первый дескриптор запроса является важнейшим по отношению к другим. Последующие дескрипторы являются уточняющими. Для проверки этого предположения был проведен эксперимент. В пяти поисковых системах (www.rambler.ru, www.yandex.ru, www.gogo. ru, www.yahoo.com, www.nigma.ru) сети Интернет введен сначала один дескриптор, потом два, три и четыре. Поисковые системы выдавали при двух дескрипторах поиска в среднем в два раза меньше документов, чем при одном поисковом дескрипторе. Такая же закономерность прослеживалась при вводе трех и четырех дескрипторов.

Алгоритм ИПС ВМН завершается выдачей отчета (результата) пользователю, используя две формы представления данных: таблицу и список. Пользователю предполагается сделать выбор формы представления найденной информации, предусматривается возможность совмещения обеих форм (таблица и список) для представления результата.

В случае, если ИПС ВМН выдала неполный отчет, пользователь может попробовать уточнить запрос и выполнить поиск повторно. Если массив выданных данных очень велик, пользователь может его уменьшить путем изменения запроса и осуществления поиска уже в найденном массиве.

Представленный алгоритм осуществляет поиск по нескольким базам данных. Использование весовых параметров привело к формализации правила составления поисковых запросов. Чем выше вес дескриптора, тем он важнее для пользователя. Алгоритм ИПС ВМН предоставляет наиболее полный отчет на запрос пользователя и является очередным шагом вперед к созданию идеальной ИПС.

Использованная литература

1. Аветисян Р.Д. Теоретические основы информатики / Р.Д. Аветисян, Д.О. Аветисян. – М.: Российск. гос. гуманит. ун-т, 1997. – 168 с.

2. Блюменау Д.И. Информационный анализ/синтез для формирования вторичного потока документов / Д.И. Блюменау. – СПб.: Профессия, 2002. – 240 с.

3. Гусев В.С. Google: эффективный поиск. Краткое руководство / В.С. Гусев. – М.: Вильямс, 2006. – 240 с.: ил.

4. Захаров В.П. Информационные системы (документальный поиск) / В.П. Захаров. – СПб., 2002. – 188 с.

5. Романенко В.Н. Сетевой информационный поиск : практ. пособ. / В.Н. Романенко, Г.В. Никитина. – СПб.: Профессия, 2005. – 288 с.

6. Шемакин Ю.И. Теоретическая інформатика : учеб. пособ. / Ю.И. Шемакин; под общей ред. К.И.Курбакова. – М.: Рос. экон. акад., 1998. – 132 с.: ил.