УДК 004.021

А.С. ГОДУНОВ

Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Украина

АНАЛИЗ МЕТОДОВ ВЫЧИСЛЕНИЯ КОЭФФИЦИЕНТА РЕЛЕВАНТНОСТИ ДЛЯ ВЕБ-СТРАНИЦ ПОИСКОВЫМИ СИСТЕМАМИ

Приведен метод расчета релевантности для веб-ресурсов, который позволяет учитывать неограниченное количество критериев, а так же имеет очень простую для программирования структуру, которая легко расширяется или сужается, в зависимости от количества используемых критериев. Аналитические методы доказывают возможность соответствия данного метода, тем методам, которые применяются в популярных поисковых системах Интернета. В данной статье рассматривается больше математическая модель и особенности её организации, нежели практическое использование данной модели на практике. На сегодняшний момент работа находиться на этапе уточнения параметров, применяемых в этой модели. Следующим этапом планируется проведение опытных испытаний по расчету релевантности веб-ресурсов на основе окончательной модели.

Ключевые слова: *релевантность, поиск по тексту, коэффициент релевантности, веб-документ.*

Введение

Спустя около двух десятков лет с момента появления веб-сайтов доступных общественности, мы можем насчитать уже далеко за 10 млн. таких ресурсов. Если быть более точным, то их кол-во оценивается во всем мире величиной в районе 300 млн. единиц (из них украинских ресурсов 0,5 млн.). Если предположить, что в среднем каждый такой ресурс имеет от 20 до 50 страниц, то в сумме мы получим от 6 до 15 млрд. веб-документов. Когда же речь заходит о поиске нужной информации в сети Интернет, то становиться очевидным, что требуются довольно сложные алгоритмы выборки нужной информации, не говоря уже о больших вычислительных мощностях, требуемых для исполнения этих алгоритмов.

В настоящий момент алгоритмы поиска релевантных текстов применяются поисковыми системами в совокупности с другими методами для расчета значимости Интернет ресурсов по различным фразам.

Сейчас в основе поиска релевантных текстов используются методы, основанные на словоформах и словообразованиях, которые позволяют для русского и украинского языков расширить точность поиска по сравнению с точностью поиска по обычному совпадению. При этом существует тенденция по расширению точности поиска от словоформ до синонимов.

Таким образом, сами алгоритмы поиска модифицированы в версии, которые направлены на решения задачи релевантного поиска по смыслу, вместо обычного поиска по совпадению. Такой подход позволяет строить интеллектуальные алгоритмы поиска, основанные на речевых особенностях каждого конкретного языка.

Сейчас поисковыми системами используется до 200 критериев в общей оценки релевантности документа (по данным самих разработчиков). Большинство из критериев рассчитываются на основе данных собранных самой поисковой системой, и при этом эти данные, как правило, не полные, а их объем ограничивается производительностью всей системы.

Исходя из этого, поисковыми системами упор делается на наиболее качественную фильтрацию собранных данных, чтобы уменьшить неточности вводимые недостатком объема этих данных.

Учитывая бурное развитие Интернета и рост количества интернет ресурсов, текущие алгоритмы оценки релевантности, используемые на введенных в эксплуатацию модулях поиска, очень быстро устаревают или теряют точность. Именно поэтому в настоящий момент частота обновления поисковыми системами алгоритмов оценки релевантности ресурсов меняется довольно часто (1-2 раза в год), что говорит об отдаленности текущих методов от идеального алгоритма.

Именно поэтому поиск и разработка новых алгоритмов не прекращаются, и, думаю, будут продолжаться еще довольно долго – как минимум до появления идеологически новых подходов в решении задач такого плана.

1. Постановка эксперимента

1.1. Суть эксперимента

Ход экспериментальных работ по сбору данных для анализа алгоритмов расчета релевантности веб-страниц, применяемых поисковыми системами, был организован на ряде веб-сайтов, которые уже имеют свой рейтинг в поисковых системах Google, Yandex и др. Суть опытных экспериментов сводилась к обнаружению максимального кол-ва критериев, которые могут влиять на релевантность ресурса в поисковой системе. Учитывая то, что структура всех веб-документов стандартная, т.е. одинаковая в большинстве случаев, то каждый конкретный опыт сводился к отслеживанию первоначального рейтинга релевантности веб-документа, дальнейшего внесения в этот документ изменений и отслеживание полученного через некоторое время нового рейтинга релевантности документа в поисковой системе. Учитывая то, что каждый опыт состоял из внесения изменений в отдельные элементы документа, то, сопоставляя начальный и полученный рейтинги, выявлялись наиболее значимые части документа. При этом отмечался сам факт влияния и вес этого влияния. Так же стоит отметить, что релевантность всегда определяется как отношение одного текста другому по смыслу. В нашем случае искомый текст – это фраза, вводимая в поисковую систему через строку запроса, а веб-документ – это текст, в котором выполняется поиск. Поэтому оценка рейтинга релевантности для одного и того же документа может выполняться по множеству фраз, что существенно упрощает само отслеживание, так как можно после каждого опыта проверять релевантность сразу по множеству фраз, тем самым, получая более объективную картину, которая исключает большое кол-во неточных выводов.

1.2. Результаты эксперимента

Эксперименты проводились на ряде сайтов, посвященных разным тематикам, чтобы избежать субъективности в получении результатов. Учитывая, что тенденции оказались общими, то в табл. 1-2 вы можете ознакомиться с выборочными результатами экспериментов, на основе одного сайта:

А) Изменение содержимого страницы (т.е. наполнение внутри тега BODY), оценка выполнялась на основе частотности фраз.

Фраза «**тренажеры**»:

Таблица 1

Результаты эксперимента по частотности фразы

№ эксперимента	Частота %	Частота позиция	Позиция в ПС
1	1,67	3	24
2	2,36	1	14
3	3,05	1	12
4	2,39	2	19

В данном примере «Частота %» – это коэффициент частотности фразы в тексте.

«Частота позиции» – это позиция в рейтинге частотности фраз на странице. Например, 1 – это самое частотное, 4 – это означает, что есть еще 3 фразы, у которых коэффициент частотности выше.

Позиция в ПС (Поисковой системе) – это позиция сайта в поисковой системе Google, получаемая при вводе тестируемого запроса.

Б) Изменение ключевых тегов с информацией о содержимом страницы, тег Title.

Фраза «**тренажеры**»:

Title 1: Магазин тренажеров и спорттоваров;

Title 2: Магазин ГТО предлагает тренажеры и спорттовары;

Title 3: Тренажеры и спорттовары в магазине ГТО.

При этом полученные позиции в ПС распределились таким образом:

Title 1 – 42,

Title 2 – 31,

Title 3 – 18.

Таким образом становиться очевидным, что данный элемент веб документа очень сильно влияет на позицию в выдаче, так как разброс позиций имеет наибольшую амплитуду, по сравнению с изменением содержимого страницы.

В) Изменение наполнения веб-страницы путем применения и удаления выделяющих тегов, таких как теги текстовых заглавий H1, тег подсветки жирным шрифтом B, STRONG.

Проведенные эксперименты показали, что при использовании «подсветки» для фразы, отличной от тестируемой, позиция сайта по этой фразе в поисковой системе не менялась. В случае же выделения тестируемого запроса, рейтинг страницы поднимался, но незначительно.

Г) Изменения текста в гиперссылках, указывающих на страницу, участвующую в рейтинге ПС.

Тестируемая фраза «**тренажеры**»:

Таблица 2

Результаты эксперимента по влиянию внешних ссылок на рейтинг ресурса в ПС

Ссылок	С фразой	Без фразы	Позиция в ПС
0	0	0	47
1	0	1	47
2	1	1	45
5	4	1	33
8	7	1	29
7	7	0	29

Исходя из данного эксперимента, можно сделать вывод, что вес странице добавляется от внешних только для тех фраз, которые присутствуют в тексте ссылки. Так же, судя по амплитуде изменения рейтинга в поисковой системе, можно сделать и другой вывод – ссылки, указывающие на страницу, являются очень весомым критерием при расчете

рейтинга релевантности поисковой системой.

Так же было проведено еще множество тестов, которые не вошли в эту статью по причине сложности вместить этот материал в требования издания. Однако результаты, которые не вошли в эту статью никак не влияют на общую идею предлагаемой методики расчета релевантности веб-ресурсов.

2. Анализ полученных результатов

На основе проведенных опытов было выявлено N критериев, которые однозначно влияли на общий рейтинг релевантности документа. Ниже приведены полученные опытным образом критерии:

1. Релевантность заглавия документа (содержимое тега TITLE);
2. Релевантность основного текста документа (очищенное содержимое тега BODY);
3. Релевантность внутренних заголовков (содержимое тега H1, H2 и т.д.);
4. Релевантность смысловых выделений в тексте (содержимое тегов B, STRONG и т.д.);
5. Релевантность краткого описания документа (META информация, именуемая description);
6. Релевантность ключевых слов документа;
7. Релевантность текста в гиперссылках, указывающих на данный ресурс, и др.

Список критериев, приведенных в этой статье, не полный (он достаточно велик, при этом другие критерии менее значимы), однако, дальнейшее рассмотрения алгоритма вполне возможно даже с этими исходными данными.

Построенный список упорядочен по значимости каждого из выявленных критериев с учетом влияния на общий рейтинг документа. Причем стоит отметить, что разные поисковые системы по-разному реагировали на данные критерии. Поэтому в ходе анализа результатов экспериментов стало ясно, что важность каждого из критериев – это, скорее всего, плавающая величина, которая не является константой. На основе таких выводов было решено ввести некую систему именований, на основе которой можно было бы изложить алгоритм расчета релевантности веб-документа в более строгом математическом виде.

2.1. Определение переменных для формулировки алгоритма

Исходя из того, что все же каждый из критериев влияет на общую релевантность документа, но при этом по разному для каждой поисковой системы, то самым логичным решением было разделить релевантность по каждому из критериев на две составляющие:

1. Математическая величина релевантности по данному критерию, полученная в ходе вычислений (будем называть её коэффициентом релевантности);
2. Вес критерия при вычислении общей релевантности документа в поисковой системе.

Таким образом, пусть первая составляющая именуется X, а вторая составляющая K, тогда релевантность по определенному критерию для конкретного документа будет выражаться как X*K.

Если считать, что математическая релевантность – это величина в диапазоне [0;1], то тогда можно вычислить общую релевантность документа как сумму релевантности критериев. При этом если установить, что коэффициент K может быть только в диапазоне [0;1], а сумма всех коэффициентов всех критериев для документа дает 1, то получается, что общая релевантность документа тоже будет выражена значением, находящимся в диапазоне [0;1].

Таким образом была получена формула для вычисления релевантности документа, которая легко может быть выражена в математической терминологии.

Переменные:

i – количество критериев для вычисления релевантности документа, всегда >= 1;

X_i – математическая величина релевантности по критерию, где $X_i = [0;1]$;

K_i – вес критерия, установленный поисковой системой, где $K_i = [0; 1]$, а $\sum K_i=1$;

R_i – коэффициент релевантности по критерию i, $R_i = [0; 1]$;

R – общий коэффициент релевантности документа по рассчитываемой фразе.

Получаем такую формулу:

$$R = \sum R_i = \sum K_i X_i, \quad (1)$$

где R всегда в диапазоне [0;1], при этом

0 – это полное несоответствие,

1 – это абсолютное соответствие, которое обычно может быть получено на практике только при расчете релевантности документа по фразе, где фраза идентична самому документу.

2.2. Повышение точности полученной формулы

Если мы попытаемся дать смысловую оценку данной формулы, то получиться что она работает по накопительному принципу, что не позволяет более жестко фильтровать ресурсы поисковыми системами. Например, если какой-то ресурс был отмечен как некачественный ресурс (спам ресурс, либо несанкционированная копия существующего веб-ресурса), то было бы очень полезно иметь средства или методы, позволяющие самой поисковой системой принудительно занижать рейтинг такого ресурса при вычислении его релевантности.

Проводя разбор формулы (1) становиться очевидным, что важно обеспечить возможность задавать K_i в диапазоне [-1;1], но при сохранении условия $\sum K_i=1$. Это предоставит возможность вводить как критерии релевантности, так и критерии штрафных санкций. Причем последние на практике существуют, и это не скрывается самими разработчиками поисковых систем. Такие критерии штрафных санк-

ций называются фильтрами, накладываемыми на веб-документ при формировании рейтинга документов по определенному запросу.

В результате проведенного анализа была получена методика вычисления релевантности веб-документов, причем эта методика имеет гибкую формулировку, что позволяет её модифицировать с целью получения более точных значений путем введения новых критериев, кроме того, она позволяет учитывать в вычислениях так называемые фильтры поисковых систем.

Учитывая что реальные алгоритмы поисковых систем закрыты, то на настоящий момент нет возможности доказать или опровергнуть со 100% точностью достоверность этой формулы, но то что эта формула имеет довольно объективное обоснование её существования, а так же легко поддающийся программированию алгоритм расчета доказывает возможность частичного или полного соответствия её реальности.

Заключение

Подводя итог всему вышеописанному, следует отметить, что пока еще нет опытных данных, которые на практике демонстрируют использование предложенной методики. Основываясь на начальных опытах, можно полагать о состоятельности алгоритма. Если же оценить применимость этого алгоритма в поисковых системах Яндекс и Google, то можно сказать что наличие весовых коэффициентов дает возможность корректировать точность вычислений этими поисковыми системами, придавая большие или меньшие веса тому или иному критерию. Тем самым одна и та же формула в разных реализациях (т.е. с разными весовыми коэффициентами и одинаковым набором критериев) на одних и тех же исходных данных может давать абсолютно разные результаты.

Литература

1. Капустин В.А. Основы поиска информации в Интернет / В.А. Капустин. – СПб., 1998. – 312 с.

2. Федоровский А.Н. / А.Н. Федоровский, М.Ю. Костин // Труды третьего российского семинара по оценке методов информационного поиска РОМИП-2005. – СПб.: НИИ ХИМИИ СПбГУ, 2005. – С. 109-124.

3. Brin Sergey The Anatomy of a large-scale hyper textual web search engine [Электрон. ресурс] / Sergey Brin, Lawrence Page. – Режим доступа к ресурсу: http://infolab.stanford.edu/~backrub/google.html.

Поступила в редакцию 1.03.2010

Рецензент: д-р техн. наук, проф., директор института С.Г. Антощук, Одесский национальный политехнический университет, Одесса, Украина.

АНАЛІЗ МЕТОДІВ ОБЧИСЛЕННЯ КОЕФІЦІЄНТА РЕЛЬОВАНТНОСТІ ДЛЯ ВЕБ-СТОРІНОК ПОШУКОВИМИ СИСТЕМАМИ

О.С. Годунов

Приведений метод розрахунку рельовантності для веб-ресурсів, який дозволяє враховувати необмежену кількість критеріїв, а так само має дуже просту для програмування структуру, яка легко розширюється або звужується, залежно від кількості використовуваних критеріїв. Аналітичні методи доводять можливість відповідності даного методу, тим методам, які використовуються в популярних пошукових системах Інтернету. У даній статті розглядається більше математична модель і особливості її організації, ніж практичне використання даної моделі. На сьогоднішній момент робота знаходитися на етапі уточнення параметрів, вживаних в цій моделі. Наступним етапом планується проведення дослідних випробувань за розрахунком рельовантності веб-ресурсів на основі остаточної моделі.

Ключові слова: рельовантность, пошук по тексту, коефіцієнт рельовантності, веб-документ.

ANALYSIS OF RELEVANCE RATIO's CALCULATION METHODS FOR WEB PAGES BY SEARCH ENGINES

O.S. Godunov

Article describes the method of calculating the relevance ratio for web pages, which could be used by search engines like Google, Yandex or any other as the top level algorithm. This method has an easy programmed body, which could be enlarged to as much criteria as needed. The analysis of this method proves that it could be used in real working search engines. It is allow taking into consideration all possible criteria with positive and negative influence. The list of criteria is also discussed in this article. Real state of algorithm development is the mathematical model, which is in the phase of collecting the criteria. The next step will be the programming of search engine core based on this algorithm.

Keywords: text relevance, text search, relevance ratio, web pages, search engines.

Годунов Александр Сергеевич – ст. преп. кафедры «Компьютерные системы и сети», Национальный аэрокосмический университет им. Н.Е. Жуковского «ХАИ», Харьков, Украина, e-mail: alex@uh.ua.