УЧЕТ ГИПЕРТЕКСТОВЫХ ССЫЛОК МЕЖДУ ДОКУМЕНТАМИ ПРИ РАНЖИРОВАНИИ РЕЗУЛЬТАТОВ ПОИСКА

Шарапов Р.В.,Шарапова Е.В., Торохова Е.А.

E-mail: info@vanta.ru

Муромский институт Владимирского государственного университета, г. Муром

Аннотация. *В работе рассматриваются вопросы улучшения качество поиска за счет учета гипертекстовых ссылок между документами. Дается понятие алгоритма PageRank. Рассматривается алгоритм Клейнберга (HITS).*

Hypertext links in documents as ranging results of search

Sharapov R., Sharapova E., Torohova E.

Abstract. *In work questions of improvement quality of search due to the account of hypertext links in documents are considered. The concept of algorithm PageRank is given. Algorithm HITS is considered.*

В ходе постоянного усложнения поисковых систем в интернет и увеличения доступных для поиска документов (поисковая система Яндекс работает с более 2 миллиардами страниц) возникает задача улучшения качества поиска за счет определения наиболее авторитетных источников и отсечения недостоверных данных. Одним из методов решения этой задачи является учет гипертекстовых ссылок между документами.

Парадигма гипертекста, или связанных документов, является фактическим стандартом современных информационных систем. Более того, эта информация широко использовалась еще в докомпьютерную эпоху: анализ цитирования широко применялся в библиотечном деле на протяжении веков. Идея о том, что связи содержат информацию, которая может быть полезна с точки зрения информационного поиска высказывалась неоднократно. При этом приводились следующие доводы [2]:

1. если документ A ссылается на документ B, то это значит, что автор документа A рекомендует документ B. Таким образом, это говорит о том, что документ, на который имеются гиперссылки, имеет более “высокое качество"или является более значимым;
2. если документы A и B связаны между собой гиперссылками, то существуют вероятность того, что эти документы на одну тему выше, чем в случае, если они не связаны.

Другой подход, который можно применить, обосновывая возможную полезность связей — это представление гипертекстовой коллекции как многосвязного графа, по которому осуществляет переходы читатель [2]. При этом он осуществляет переход из одного узла графа в другой с некоторой вероятностью, а процесс его навигации можно описать Марковской вероятностной моделью. В этой модели вся коллекция представляется как матрица вероятностей перехода в ту или иною вершину графа — документ. При этом можно определить, какие документы будут значительно вероятнее посещаться, а значит, и будут более значимыми в данной коллекции. Однако первые исследования не показали заметного улучшения качества поиска при использовании информации о ссылках. Это объяснялось отчасти тем, что исследователи использовали в качестве экспериментальных коллекций некоторые области сети Интернет, которые не содержали полных гипертекстовых графов. Настоящий прорыв в данной области произошел, когда поисковая система Google стала одним из самых популярных поисковых систем сети Интернет. При этом ее разработчики широко использовали анализ гипертекста и позиционировали эти возможности как одно из основных преимуществ системы [1]. Их успех вдохновил множество команд на исследования в данной области. Сейчас трудно назвать систему, которая не использовала бы информацию о гипертексте для поиска.

Модель, основанная на “Индексе цитирования”, или PageRank. Данная модель является одной из простейших, учитывающих гипертекст. В данном случае к модели документа, основанной на терминах документа, добавляется только одно число — ранг документа, вычисляемый на основании гипертекстовых ссылок в системе. С точки зрения приведенных в начале данного раздела предположений, PageRank является вероятностью того, что “случайный читатель“ посетит данный документ или, с точки зрения другого подхода, что данный документ имеет определенное “качество“, содержащейся в нем информации. Классическая формула вычисления PageRank следующая [1]:

Предположим, страница A имеет страницы T1... Tn, которые на нее ссылаются. (A) — количество ссылок, которые выходят из страницы A. Тогда для данной страницы формула PageRank будет выглядеть следующим образом:

$PR(A) = (1-d) + d(PR(T1))/C(T1) + ... + d(PR(Tn))/C(Tn)$

где d - некоторый коэффициент, который выбирают обычно равным 0.8-0.9.

Как видно, формула является итерационной, так как ранг данной страницы зависит от ранга других страниц, ранг которых, в свою очередь, может зависеть от данной. Существует большое количество работ, посвященных вопросам эффективного вычисления PageRank, особенно в случае очень больших графов, каким является современный Интернет. Обычно для этого применяют различные методы экстраполяции, позволяющие снизить количество итераций при вычислении. Данная модель имеет ряд преимуществ, к которым можно отнести следующее:

1. простота реализации. Для каждого документа добавляется всего один коэффициент;
2. малые затраты при поиске и относительно малые затраты при индексировании.

PageRank очень популярен в современных коммерческих системах и используется практически всеми общедоступными поисковыми Интернет-машинами. Ее успех во многом основывается на том, что ее достаточно трудно фальсифицировать, создав специальные страницы (так называемая поисковая оптимизация сайтов). Однако, в последнее время фирмы, специализирующиеся на “вытаскивании"определенных сайтов в начала списков по запросам в популярных поисковых машинах, научились “обманывать” PageRank, создавая специальные страницы, содержащие много ссылок. В целях борьбы с этим явлением, и увеличения качества поиска система Google в конце 2003 года применила еще один дополнительный параметр, который можно рассматривать как достаточно простую модификацию PageRank — так называемый HillTop [2]. В нем при вычислении PR(A) используются не все ссылки на данную страницу, а только ссылки с определенного подмножества сайтов, выбранных экспертами как достоверные. При этом в модели документа хранится не один коэффициент PageRank, а два — вычисленный по всей коллекции и только по “достоверным"документам. Данный коэффициент позволяет оценить, применялись ли специальные технологии искажения PageRank для данных страниц и провести необходимую коррекцию результата. По данным Google, приведенным в упомянутой статье, этот алгоритм позволяет увеличить на несколько процентов полноту и точность поиска.

Модели, учитывающие контекст запроса. Описанный коэффициент, основанный на индексе цитирования, вычисляется статически для документа, при этом не учитывается тема документа. Например, если документ имеет высокий PageRank, но не относится к данной теме, то он будет с точки зрения пользователя переоценен системой. Учет информации о запросе при анализе гипертекстовых ссылок должен компенсировать этот дефект. Существует несколько подходов, которые позволяют это сделать, но в основе всех этих подходов лежит одна идея: ссылки анализируются не по массиву всех документов, а по множеству документов, отобранных в качестве релевантные данному запросу. В простейшем случае модель документа расширяется массивом коэффициентов PageRank, вычисленных на подмножестве документов, которые относятся к некоторому фиксированному набору тем [3].

Достоинством такого подхода при ограниченном числе тем является то, что вычисления выполняются не в момент поиска, а являются предварительным этапом, что позволяет выполнять поиск достаточно быстро, при этом затраты на вычисления массива PageRank незначительно больше, чем вычисление коэффициента, не связанного с темой. Такое расширение модели документа не приводит к значительному росту требуемой памяти, т.к. тем, для которых рассчитывается ранг, немного. Недостатком такого метода является то, что он требует отнесения запроса к одной из тем, что не всегда возможно и может вносить дополнительную погрешность. Авторы данной методики в указанной работе провели исследования, которые показали, что применение этого подхода позволило получить значительно более высокую точность поиска.

Однако на практике выделение тем документов является отдельной сложной проблемой, и применяются методы, учитывающие гиперсвязи на этапе обработки запроса. При этом модель документа должна быть расширена дополнительной информацией, которая представляет собой список идентификаторов документов, на которые ссылается данный. Хранение этой информации может потребовать достаточно много памяти и вычислительных ресурсов. Поэтому данные методы только начинают применяться в современных поисковых системах.

Самым простым из данных подходов является LocalRank. Данный алгоритм является вариантом PageRank, который вычисляется не на всей коллекции, а только на документах, которые отобраны из коллекции при обработке запроса другими методами.

Другой метод учета, напротив, впервые был предложен теоретически и имеет достаточно обширную библиографию исследовательских работ, но не нашел широкого применения в коммерческих системах. Это алгоритм, который получил название HITS, или алгоритм Клейнберга [4]. В его основе лежит предположение, что все документы можно разделить с точки зрения гипертекстового графа на два типа: авторитетные источники (authority) — документы, содержащие важную информацию, на которые ссылаются другие документы, и документы-узлы (hub), которые содержат ссылки на документы-источники информации. Считая, что на хороший источник информации ссылаются хорошие документы-центры, получаем список хороших документов-центров. Далее эти документы-центры ссылаются на хорошие источники и т.д. В результате получается следующий рекурсивный алгоритм:

1. пусть N — множество документов, отобранных запросом, при том, что для каждого нам известны ссылки на другие документы запроса;

2. для каждого документа n из N создаем два массива задаем два значения: H[n] — оценка документа-центра, инициализируем это значение 1, и A[n] — оценка документа-источника;
3. далее выполняем в цикле следующие два действия пока значения оценок A[n] и H[n] не перестанут изменяться с заданной точностью (т.е. пока не придут к установившимся значениям);
4. для каждого из документов вычислить A[n] как сумму H[k], где k - множество документов, которые ссылаются на данный;
5. для каждого из документов вычислить H[n] как сумму A [к], где k - множество документов, которые ссылаются на данный;
6. нормализуются значения вычисленных A[n] и H[n] на множестве документов.

Документы, которые получили наибольшие значения A[n] и H[n] имеют больший вес в результате поиска.

Модель, "переносящая" термины. Данная модель описана в работе [1] и интересна тем, что представляет собой модификацию модели документа "множества весов слов", но учитывает гипертекстовые ссылки. При этом подходе к множеству терминов документа добавляют множество терминов, находящихся в окрестностях гиперссылок на данный документ. Подход дает хороший практический результат, когда коллекция содержит документы, которые имеют мало текста — например, документы с графическими изображениями или записями звуков.

Качество поиска с использованием моделей, учитывающих связи документов. Как уже говорилось, первые исследования таких алгоритмов не показали значительного улучшения результатов. В то же время они стали широко применяться в коммерческих системах поиска по Интернет. Это объясняется тем, что учет связей документов позволяет эффективно выделять ресурсы высокого качества и отсеивать фальсифицированные ресурсы.

Методы, учитывающие контекст запроса, значительно лучше исследованы. Это можно объяснить тем, что на них менее влияет то, что выбираемая для исследования коллекция является подграфом, т.е. большое количество ссылок направлено вне коллекции. Большинство работ показывают рост точности поиска до 50%, полнота при этом обычно не меняется, ведь данные алгоритмы не добавляют новые документы в результат поиска. При этом общей проблемой всех моделей, учитывающих контекст, является так называемый "дрейф" [4]. Это явление, которое возникает, когда в результат запроса, обрабатываемый таким алгоритмом, попадает группа документов, хотя и мало релевантных, но сильно связанных между собой. При этом эта группа оценивается значительно выше, что приводит к смещению результата от темы запроса, к теме группы связанных документов. Для нейтрализации этого явления предложен ряд оригинальных подходов, в основе которых лежит учет не только факта наличия связи на документы, но и вес документа, на который указывает связь и контекст ссылки.

В современных коммерческих системах методы, учитывающие контекст запроса, применяются не очень широко. Поисковые машины в Интернет не используют данный подход, так как он требует значительных ресурсов на этапе обработки запросов, что достаточно критично для этого вида приложений. Кроме того, в отличии от PageRank, при небольшом количестве документов, содержащих термины запроса, данные показатели намного проще фальсифицировать, создав специальные страницы. Однако в последних версиях крупнейшей поисковой системы Google заявлено использование алгоритма Local PageRank и даже запатентована его реализация, что говорит о том, что в ближайшем будущем мы увидим более широкое применение этого подхода, потенциально имеющего большие возможности [2].

Качество информационного поиска современных систем неуклонно растет. Это обусловлено растущими вычислительными мощностями и объемами памяти современных компьютеров, что позволяет реализовывать все более сложные и ресурсоемкие алгоритмы. Многие описанные модели, особенно связанные со сложным анализом текста, такие как синтаксический анализ, пока просто невозможно реализовать в полном объеме, поэтому трудно оценить их влияние на качество поиска. Другие пока еще недостаточно изучены с точки зрения их влияния на качество поиска и требуют дополнительных исследований.

Нельзя сказать, что существующие модели представления документа позволяют использовать для информационного поиска всю информацию, которую можно извлечь из документа. Очевидно, что можно создать новые и/или использовать комбинации уже существующих моделей.

Литература

1. Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1_7):107-117, 1998.
2. Губин М.В. Модели и методы представления текстового документа в системах информационного поиска: Диссертационная работа к.т.н.: 05.13.11 / Санкт-Петербургский государственный университет – СПб., 2000. – 95 с.
3. T. Haveliwala. Topic-sensitive pagerank. In Proceedings of the Eleventh International World Wide Web Conference, 2002.

4. Monika Henzinger. Link analysis in web information retrieval. IEEE Data Engineering. Bulletin, 23(??):3-8, 2000.