Ю. В. Малинина

ПРОГРАММНЫЕ СРЕДСТВА ДЛЯ АВТОМАТИЧЕСКОГО ФОРМИРОВАНИЯ ТЕМАТИЧЕСКОЙ КОЛЛЕКЦИИ ПО ПРЕОБРАЗОВАНИЯМ ПРОГРАММ ДЛЯ КОЛЛЕКТИВНОГО ИСПОЛЬЗОВАНИЯ*

ВВЕДЕНИЕ

В настоящее время происходят революционные процессы в изменении структуры мировой системы научных коммуникаций. В связи с этим можно отметить реферативный журнал, созданный институтом информации Гарфильда “Science Citation Index”, ориентированный на поиск новых научных публикаций в мировой системе периодических и продолжающихся изданий по системе научных ссылок. Это издание использует естественную, исторически сложившуюся систему классификации научных работ по ссылкам автора на работы его предшественников. В нем отсутствует разбиение статей на тематические рубрики, авторы указателя предлагают пользователям подготовленные ими тематические кластеры связанных между собой публикаций по системе цитирования общих предшественников. Как размеры, так и наименования кластеров постоянно корректируются ими в соответствии с тенденциями в науке.

Следующим шагом в развитии коммуникации в мировом научном сообществе стало широкое распространение электронной почты, позволившей многим ученым реализовать каналы неформальной коммуникации, которая ранее могла происходить только на конференциях, симпозиумах и семинарах. Это привело к образованию большого числа “невидимых колледжей”, неформальных объединений ученых, работающих в одной тематической области науки. Этот процесс неформальной коммуникации, происходящий в настоящее время, еще достаточно подробно не изучен и ждет своих исследователей.

Наиболее значимым событием в развитии системы научной коммуникации явилось появление в Internet (мировой информационной сети) информационных страниц различных научных школ (университетов, научных

* Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (грант № 01-01-794) и Министерства образования РФ.

институтов, коллективов и т. д.) и даже отдельных ученых. Одним из важнейших событий стало представление в них библиографий научных документов (публикаций в мировой научной печати, сообщений, тезисов и т. д.), созданных в рамках их научных исследований. Многие ученые поняли важность такой информационной работы, позволяющей сохранить для следующих поколений исследователей результаты их научной работы.

Наиболее ценными информационными объектами среди них можно считать списки полных библиографий научных документов, относящихся к одной тематической области, которые создавались энтузиастами-учеными и целыми международными коллективами в рамках исследовательских проектов.

Задача автоматического создания и дальнейшего пополнения подобной тематической коллекции в области преобразований программ рассматривается в данной статье.

СБОР ИНФОРМАЦИИ

Для получения информации в Интернете в основном используются поисковые системы общего назначения, которые выполняют индексирование всех существующих страниц. Построение индексов Интернет-ресурсов для поисковых систем традиционно основано на использовании сетевых роботов-программ, которые, начиная с некоторой Интернет страницы, рекурсивно обходят ресурсы Интернет, извлекая ссылки на новые ресурсы из получаемых документов. В каждом случае используется некоторый определенный метод извлечения информации.

Особенности задачи поиска в Интернете

Методы поиска, используемые в классических поисковых системах, разрабатывались и тестировались на относительно небольших, однородных коллекциях документов, например, таких как библиотечные каталоги или коллекции газетных статей. Интернет как набор данных имеет ряд важных особенностей.

1. Данные в Интернете организованы крайне стихийно и не систематично. Несмотря на то что принято считать, что Интернет — это распределенный гипертекст, это не совсем так. Гипертекст обычно подразумевает наличие концептуальной модели, которая накладывает ограничения согласованности на данные и гиперсвязи. В Интернете это обычно не так

даже для тех его частей, которые находятся под единым административным контролем. Около 30% информации в Интернете составляют точные или приблизительные копии других документов.

2. Текущий объем доступной информации в Интернете оценивается в десятки терабайт и быстро возрастает, поэтому даже самый мощный сетевой робот не может отслеживать содержание всех ресурсов Интернета. Отметим, что эти оценки касаются только той «поверхностной» части Интернета, которая не скрыта за поисковыми формами и доступ к которой не требует предварительной регистрации или авторизации. Другую, «скрытую», часть Интернета (hidden web) поисковые системы обычно не рассматривают, а ведь к ней относится множество крупных баз данных, опубликованных в Интернете. Поэтому неудивительно, что оценка объема «скрытого» Интернета в 500 раз больше, чем объем «поверхностного» Интернета.

3. По некоторым оценкам ежемесячно публикуется около 30 миллионов новых документов, причем ежемесячно изменяется до 40% ранее опубликованной информации. Среднее время жизни Интернет-страницы около 24 дней.

4. Большая доля документов виртуальна в том смысле, что формируется в ответ на некоторый запрос пользователя и не хранится в явном виде. Как правило, это результаты поиска в различных базах данных.

5. В Интернете используются более 100 естественных языков. Сами документы представляются в различных форматах, таких как html, xml, Word, PostScript, PDF и т.п.

6. Отсутствие редакторского контроля над публикуемой информацией в Интернете обуславливает проблему с ее качеством: информация может быть некорректной (например, уже устаревшей), ложной, плохо сформулированной, содержать массу ошибок (опечаток, грамматических ошибок, ошибок оцифровки и т.п.). Так, по некоторым оценкам, одна опечатка встречается в среднем в каждых двухстах часто употребительных словах или в трех иностранных фамилиях.

Особенности структуры Интернета

Интуитивно кажется естественным предположение о том, что ссылки с некоторой Интернет-страницы в основном ведут на страницы близкой тематики. Эмпирическое подтверждение согласованности между тематической и пространственной (в смысле расстояний в графе Интернет-страниц) локальностью было дано в работе [4].

«Тематическое Сообщество» можно определить как совокупность страниц, каждая из которых имеет больше ссылок (в любом направлении) внутри этой совокупности, чем снаружи. Данное определение можно обобщать, для того чтобы выявить объединения различного размера с различным уровнем связности. Более строгое определение и алгоритм выявления таких сообществ можно найти в [6].

Подходы к организации поиска

Существуют два основных подхода. Первый из них заключается в использовании индексов существующих универсальных поисковых систем. Этот подход достаточно широко применяется на практике и имеет свои как положительные, так и отрицательные стороны.

К его положительным сторонам относятся:

- повторное использование ранее полученных данных. Сканирование Интернета — это дорогостоящий процесс, который приводит не только к большим затратам для проводящей его организации, но и затрагивает другие интересы владельцев индексируемых сайтов, пользователей. Представляется весьма неразумным многократно сканировать Интернет ради предоставления доступа к одной и той же информации из большого числа конкурирующих поисковых систем;
- новизна используемой информации. Универсальные поисковые системы стремятся (в идеале) индексировать все новые документы, не ограничивая себя фиксированной и, возможно, уже устаревшей тематикой. Индекс таких систем предоставляет представительную выборку относительно недавно опубликованных документов, что можно использовать при анализе тенденций в той или иной области;
- низкая стоимость получения информации. Фактически на данный момент получение информации в виде ответов на автоматически

генерируемые запросы от коммерческих поисковых систем бесплатно.

Отрицательные стороны представлены как:

- старение индекса;
- закрытость методики получения используемой информации. Алгоритмы сканирования Интернета и поиска в индексе являются коммерческой тайной, и, следовательно, индекс коммерческих универсальных систем представляется используемым его системам следующего уровня в виде черного ящика;
- невозможность делать какие-то объективные выводы о характере распределения информации в Интернете на основе косвенного (через систему поиска) анализа индекса коммерческой системы;
- ненадежность доступа к информации. Коммерческие поисковые системы очевидно не заинтересованы в том, чтобы их индекс анализировался какими бы то ни было автоматическими системами. В любой момент эксперименты в этой области могут быть запрещены в связи с нарушением тех или иных прав коммерческих поисковых систем.

Второй подход к реализации поиска связан с самостоятельным обходом Интернета и основан на использовании ссылок на новые документы из ранее загруженных документов.

Традиционно основой информационных систем является сетевой робот — это программа, которая, начиная с некоторой Интернет-страницы, рекурсивно обходит ресурсы Интернета, извлекая ссылки на новые ресурсы из получаемых документов. Для этого процесса используется метафора «ползанья» паука применительно к созданию гиперссылок в Интернете. Этот процесс повторяется с каждым новым набором страниц и продолжается до тех пор, пока не перестанут появляться новые страницы либо пока не будет собрано предопределенное количество страниц.

Фактически большинство сетевых роботов не может посещать все доступные в Интернете ресурсы из-за ограниченности доступных роботу аппаратных и сетевых ресурсов, и то, какие именно ресурсы будут посещены, определяется применяемой стратегией посещения. Естественно, что робот должен стараться использовать такую стратегию, которая максимизирует общую «полезность» всех посещенных ресурсов. Все обнаруженные, но не просмотренные страницы помещаются роботом в приоритетную очередь, упорядоченную по качеству. «Полезность» ресурса определяется той целью, для достижения которой используется робот.

Робот, который собирает информацию о ресурсах для поисковой системы общего пользования, заинтересован в обнаружении максимального количества разнообразных ресурсов. Подобные роботы зачастую используют в качестве оценки «полезности» ресурса глубину URL, т. е. количество промежуточных каталогов, упоминающихся в URL между именем Интернет-узла и именем самого ресурса. Чем больше глубина, тем ниже важность соответствующего ресурса. Такой подход позволяет быстро посетить стартовые и близкие к ним страницы на большом числе Интернет-узлов.

Остановимся на положительных сторонах второго подхода, к которым относятся:

- объективность. В данном случае информация извлекается непосредственно из сети, что обеспечивает ее объективность;
- управляемость процесса получения информации. В отличие от косвенного доступа к индексу коммерческой системы через формирование специальных запросов, доступ к информации в данном подходе более прозрачен. Имеется прямая и легко обнаруживаемая связь между алгоритмом сканирования Интернет и тематической направленностью загружаемых документов. Это позволяет подбирать параметры алгоритма, минимизирующие среднее отклонение тематики загружаемых документов от заданного тематического направления.

Отметим и отрицательную сторону второго подхода:

- высокая стоимость. В данном случае сетевой робот реально сканирует Интернет, что приводит к большим затратам даже при использовании ограниченного (заданной тематикой) поиска.

Однако давно замечено, что эффективность одного и того же метода поиска часто варьируется при применении его в различных коллекциях. Сопоставление таких наблюдений и характеристик коллекций зачастую позволяет выявить слабые места и, как следствие, способствует повышению эффективности методов поиска.

Естественным развитием этой идеи является определение характеристик коллекций документов с целью применения этой информации для повышения эффективности поиска. Это свойство позволяет рассматривать задачу автоматического формирования тематической коллекции как перспективное направление.

За последние несколько лет этой проблеме было посвящено много внимания. В данной работе рассматриваются подходы к автоматическому построению тематической коллекции в области преобразований программ.

СОСТАВЛЕНИЕ ТЕМАТИЧЕСКОЙ КОЛЛЕКЦИИ.

Рассмотрим задачу формирования тематической коллекции. Традиционно считается, что пользователь использует систему информационного поиска для обнаружения документов, содержащих информацию на связанную с его запросом тему. Однако такое предположение верно не всегда. Поскольку представление пользователя о том, что такое релевантный документ, напрямую зависит от цели, для достижения которой он проводит поиск, то естественной кажется идея оптимизировать метод поиска под конкретную цель пользователя.

В нашем случае преследуются следующие цели.

Выявление сообществ

Тематическая коллекция документов, ориентированная на определенную предметную область и формируемая на основе общедоступного Интернета, должна использовать механизмы поиска, ориентированные на определенную предметную область, и иметь возможность определять подмножества Интернета в рамках своей предметной области. Главной задачей применяемой стратегии посещения документов является выбор такого порядка обхода известных роботу ресурсов, при котором за минимальное время будет обнаружено максимальное число документов, релевантных тематике.

Составление обзора

При составлении обзора пользователю недостаточно просто найти документы с информацией по соответствующей теме. Для адекватного представления необходимо, чтобы тематическая коллекция содержала информацию с разными точками зрения. Если не учитывать специфику этой задачи поиска, то вероятно, что обнаруженные документы будут отражать только одну, доминирующую в ресурсах Интернет, точку зрения.

Один из возможных подходов к решению этой задачи состоит в построении по исходному запросу и множеству возвращаемых документов так называемого «обратного» запроса, ориентированного на получение списка документов, отражающих другие точки зрения. Для этого в исходный запрос могут быть добавлены ссылки на экспертов, которые отражают еще не найденные точки зрения, или наоборот исключены/запрещены ссылки на уже представленных экспертов. Информация об экспертах, выражающих еще не обнаруженные точки зрения по этому вопросу, может быть почерп-

нута, например, из специализированного тезауруса. Отметим, что здесь термин «эксперт» вовсе необязательно означает «человек».

Поиск по категории

Еще одним типичным примером изменения цели поиска является сужение области поиска на документы определенной категории, такой как, например, множество домашних страниц специалистов в области преобразований программ или множество анонсов научных конференций. Прямолинейным подходом к решению этой проблемы является поиск в соответствующем разделе составленного вручную каталога типа Yahoo!, но соответствующий искомой категории раздел не всегда существует, да и с большой вероятностью он содержит ссылки на далеко не все ресурсы этой категории.

Одним из возможных подходов к реализации поиска по категории является выявление различных атрибутов, характеризующих страницы данной категории, и использование этой информации для расширения запроса. Эти атрибуты далеко не всегда очевидны: так, например, при поиске по категории англоязычных домашних страниц наивное расширение запроса путем добавления «home page» повышает точность на 20%, а менее очевидные, но автоматически построенные расширения «my» и «welcome» повышают точность на 65%.

Еще один вариант — фильтрация результатов поиска. Наиболее очевидной его реализацией является применение методов бинарной классификации для определения, какие из найденных документов относятся к заданной категории.

Поиск научных публикаций

В Интернете доступно огромное количество научных работ, многие из которых еще не доступны в печатном виде, и это мотивирует интерес к задаче поиска по научной литературе.

Конечно, в некотором роде эта цель является частным случаем цели поиска по категории (см. выше). Однако этот случай выделен по нескольким причинам. Во-первых, научные статьи зачастую хранятся в форматах отличных от html. Во-вторых, поскольку научная литература гораздо больше похожа на публикации в том смысле, как этот термин понимается в библиотеках, то здесь возможно применение подходов из библиотечной области. Одним из очень полезных механизмов является использование «индекса цитирования», т.е. количества библиографических ссылок на данную работу из других статей.

МЕТРИКИ

Качество работы агента и скорость нахождения важных ссылок определяются качеством используемых фильтров. Несмотря на подтверждение того, что документы в Интернете имеют тематическую локальность для многих общих тематических областей, нет ясного понимания того, как это использовать для конкретных условий. Кроме того, в чистом виде этот подход не использует большую часть доступной информации как, например, точное содержание не просмотренных Интернет-страниц или структуру URL-кандидата. Такие данные могут обеспечить ценную информацию, для того чтобы направлять обход более эффективно. Можно ожидать, что в общем случае один из этих показателей может оказаться более важным и что упорядочение на основании дополнительных показателей может существенно корректировать направление обхода. Поэтому кажется целесообразным использовать некоторую совокупность показателей.

Примем следующие соглашения: набор Интернет-страниц рассматривается в виде направленного графа G, где каждая Интернет-страница в наборе моделируется узлом графа; если страница p_1 содержит гиперссылку на страницу p_2, то в графе есть направленное ребро (p_1, p_2); если p_1 не имеет гиперссылки на p_2, то направленного ребра (p_1, p_2) не существует.

Метрика цитируемости PR(p)

Простейшая идея глобального (т.е. статического) учета ссылочной популярности состоит в подсчете числа ссылок, указывающих на страницу. Это примерно то, что в традиционной библиографии называют индексом цитирования.

Первое допущение методов на основе связей порождает простой критерий: чем больше гиперссылок указывает на страницу, тем лучше эта страница. Основной недостаток такого подхода состоит в том, что он не делает различий между качеством страницы, на которую указывает несколько страниц низкого качества, и качеством страницы, на которую указывает то же число страниц высокого качества. Очевидно, что рейтинг страницы можно увеличить, просто создав множество других страниц, ссылающихся на данную страницу. Поэтому обычно используется модификация этой метрики, предложенная Брин и Пейдж как алгоритм PageRank [3, 4]. Он вычисляет коэффициент PageRank каждой страницы, присваивая каждой ссылке на страницу весовой коэффициент, пропорциональный качеству страницы, содержащей гиперссылку. Чтобы определить качество ссылаю-

щейся страницы, используются ее коэффициенты PageRank рекурсивно, причем первоначальные значения PageRank задаются произвольно. Точнее, $PR(p_1)$ — коэффициент PageRank страницы p_1 можно определить как

$$PR(p_1) = \frac{a}{n} + (1-a)\sum \frac{PR(p_{2)}}{outegree(p_2)},$$

где a — константа, значение которой находится в пределах от 0,1 до 0,2; n — число вершин, т.е. число Интернет-страниц в наборе; outdegree(p_2) — число ребер из p_2, т.е. число гиперссылок.

Эта формула показывает, что коэффициент PageRank страницы A зависит от PageRank страницы B, указывающей на A. Поскольку определение PageRank порождает такое линейное уравнение для каждой страницы, чтобы вычислить PageRank для всех страниц, необходимо решить огромное количество линейных уравнений. PageRank позволяет эффективно отличить высококачественные страницы Интернета от низкокачественных.

Метрика концентрации и авторитетности HR(p), AR(p)

Если предположить, что информация по теме может распределиться примерно поровну между страницами с хорошим информационным наполнением по теме, называемыми «авторитетами» (authority), и страницами, напоминающими каталоги, с множеством ссылок на другие страницы, посвященные данной теме, называемыми «концентраторами» (hub), то алгоритм Кляйнберга — поиск документов по заданной теме на базе гиперссылок (Hyperlink-Induced Topic Search — HITS) — пытается выявить хорошие концентраторы и авторитеты [7, 8]. Алгоритм итеративно вычисляет показатель концентрации и авторитетности для каждого узла графа соседей, а затем упорядочивает узлы в соответствии с этими показателями. Узлы, имеющие высокие показатели авторитетности, должны быть хорошими авторитетами, а узлы с высокими показателями концентрации должны быть хорошими концентраторами. Алгоритм исходит из того, что документ, ссылающийся на большое число других документов, — хороший концентратор, а документ, на который указывает множество других документов, — хороший авторитет. Рекурсивно документ, который указывает на большое число хороших авторитетов, — еще лучший концентратор, а документ, на который ссылается множество хороших концентраторов, — еще лучший авторитет.

Заметим, что алгоритм не утверждает, что будут найдены все высококачественные страницы, удовлетворяющие условиям запроса, поскольку некоторые из таких страниц могут не принадлежать графу соседей или входить в его состав, но не иметь ссылок со многих страниц.

Дополнительно следует учитывать, что в связи с алгоритмом HITS возникает две проблемы.

1. Поскольку рассматривается относительно небольшая часть графа Интернета, добавление ребер к нескольким узлам может серьезно изменить показатели концентраторов и авторитетов [10]. В силу этого манипулировать этими показателями достаточно просто, поэтому манипуляция ранжированием механизма поиска — серьезная проблема для Интернета.

2. Если большая часть страниц в графе соседей относится к теме, которая отличается от темы запроса, авторитеты и концентраторы, получившие высокий рейтинг, могут относиться к другой теме. Эта проблема называется «смещением темы». Добавление весов к ребрам с учетом текста документов или их тезисов значительно смягчает негативный эффект этой проблемы.

Метрика местоположения LR(p)

Метрика местоположения отражает зависимость значения страницы от ее местонахождения. Если URL заканчивается на ".com", то страница может считаться более полезной, чем URLs с другими окончаниями, или URLs содержит вхождение "home", то она может быть более интересна, чем другие URLs.

Метрика тематического соответствия IR(p, F)

Рассматривается многомерное векторное пространство, где каждому термину соответствует свое измерение. Тематический фильтр представляет собой *вектор значимостей терминов фильтра* $(F_1,...,F_n)$. Каждый документ p представлен *вектором значимостей терминов* $(p_1,...,p_n)$ в этом векторном пространстве. Формула для вычисления значимости термина (p) в документе с учетом косинусного фактора нормализации представляется формулой

$$p_i = \frac{tf_i \times idf_i}{\sqrt{\sum_i p_i^2}},$$

где tf_i (term frequency) — частота, с которой встречается данный индексный термин; idf_i (inverted document frequency) — величина, обратная частоте, с которой данный термин встречается во всей совокупности документов.

Вхождения терминов, встречающихся в документе, положительны: $p_i > 0$, а вхождения терминов, не встречающихся в документе, равны нулю: $p_i = 0$. Величина p_i представляет собой функцию, которая растет в зависимости от частоты употребления термина в документе и убывает в зависимости от числа документов в наборе, содержащем этот термин. Идея состоит в том, что чем больше документов, в которых присутствует данный термин, тем в меньшей степени термин характеризует данный документ, и чем чаще термин встречается в документе, тем в большей степени он характеризует документ.

Адекватность документа фильтру (релевантность) вычисляется как скалярное произведение их векторов терминов. В качестве результата документы упорядочиваются в порядке убывания их показателей

$$IR(p, F) = \sum p_i - F_i.$$

Многие авторы Интернет-страниц заинтересованы в том, чтобы их страницы в ответах на определенные запросы имели высокий рейтинг. Таким образом, чтобы увеличить рейтинг, авторы будут различным образом «подкручивать» свои страницы. Эти попытки манипулирования алгоритмом ранжирования иногда доходят вплоть до того, что добавляются фрагменты текста, набранные невидимым шрифтом. Например, если для ранжирования используется модель векторного пространства, добавление на страницу 1000 слов «машина» увеличит рейтинг данной страницы при поиске по запросу «машина».

Любой алгоритм, базирующийся исключительно на содержимом страницы, восприимчив к такого рода манипуляциям. Действенность анализа гиперссылок проявляется в том, что он использует для определения рейтинга текущей страницы информационное наполнение других страниц. Если предположить, что эти страницы были созданы авторами независимо от автора исходной страницы, то объективность упорядочивания при таком подходе растет.

СОСТАВЛЕНИЕ КОЛЛЕКЦИИ

Задача агента, предлагаемого в данной работе, состоит в пополнении коллекции новыми релевантными ее тематике документами. Используя описанные выше метрики, составление коллекции будет включать в себя следующие основные этапы.

1. Генерация фильтра коллекции.

Фильтр коллекции строится на основе анализа содержимого ядра коллекции. В момент анализа для всех слов из словаря коллекции (слова, встречающиеся в ее ядре) вычислялись веса. В качестве фильтра коллекции выбирается заданное количество слов из словаря коллекции с наибольшими весами.

2. Инициализация дерева URL.

На этапе инициализации формируется дерево с двумя уровнями. На первом уровне — корень дерева, на втором — узлы, содержащие стартовые URL, заданные администратором коллекции. Каждому узлу v приписывается оценка $P(v)$ вероятности того, что ссылка из соответствующего документа указывает на документ релевантный тематике коллекции

$$\mathrm{P(v)} = \frac{1}{4}(\mathrm{HR(v)} + \mathrm{AR(v)} + \mathrm{PR(v)} + \mathrm{LR(v)} + \mathrm{IR(v, F)}).$$

Для стартовых URLs на этапе инициализации эти оценки принимаются равными 1.

3. Выбор URL (этот пункт и последующие повторяются в течение всего времени работы агента).

В дереве URL выбирается узел v, для которого $P(v)$ максимально (при этом не выбираются URLs, указывающие на недавно посещенные сайты). Если документ с соответствующим URL еще не загружен, то для последующей загрузки выбирается данный URL. В противном случае, случайным образом выбирается еще не рассмотренная ссылка из этого документа на новый документ в формате html. Если нерассмотренных ссылок нет, то выбирается другой узел.

4. Загрузка и фильтрация документа.

С помощью программы wget загружается документ с заданным URL. Выполняется разбор текста документа, в процессе которого выделяются ссылки на другие html-документы, и вычисляется *tf*-вектор загруженного документа.

5. Модификация дерева URL.

Если документ не прошел фильтрацию, то он не рекомендуется для включения в коллекцию. Если для URL данного документа в дереве URL уже имеется узел, то он помечается как нерелевантный и оценка вероятности релевантности исходящих из него ссылок принимается равной нулю. В противном случае, оценка $P(v)$ уменьшается.

Если документ прошел фильтрацию, то он рекомендуется для включения в коллекцию. Если для URL данного документа в дереве URL уже имеется узел, то он помечается как релевантный и оценка вероятности релевантности исходящих из него ссылок принимается равной 1. В противном случае, оценка $P(v)$ увеличивается (если оно было меньше 1). Кроме того, формируется новый узел w, содержащий URL загруженного документа. Величина $P(v)$ принимается равной 1. В дереве URL узел w является сыном узла v.

Новое значение $P(v)$ равно

$$P(v) = \frac{1 + links_{+}(v)}{1 + links_{+}(v) + links\_(v)},$$

где $links_{+}(v)$ равно числу проверенных релевантных ссылок из документа в узле v, а $links\_(v)$ — числу проверенных нерелевантных ссылок из того же документа.

ЗАКЛЮЧЕНИЕ

Хотя на данный момент эксперименты не позволяют делать статистически значимых выводов, но в дальнейших исследованиях предполагается получить улучшенную стратегию формирования тематической коллекции.

Механизмы поиска позволяют быстро и просто получить доступ к огромным объемам информации. Их вклад в развитие Интернета и общества в целом трудно переоценить. Однако «универсальная» модель поиска в Интернете зачастую ограничивает разнообразие и полезность получаемых результатов. Предотвратить это может более активное использование контекста и тематической направленности при поиске в Интернете.

СПИСОК ЛИТЕРАТУРЫ

1. **Bharat K., Henzinger M.R.** Improved algorithms for topic distillation in a hyperlinked environment // Research and Development in Information Retrieval: Proc. / SIGIR'98: 21st Annual Internat. ACM SIGIR Conf., Melbourne, Australia, August, 1998. — ACM, 1998. — P. 104–111
2. **Chakrabarti S., Berg M. van den, Dom D.** Focused crawling: A new approach to topic-specific Interrnet resource discovery // Searching and Querying: Proc. / 8th World Wide Web Conf., Toronto, May 1999 (http://www8.org/w8-papers/5a-search-query/crawling/index.html).
3. **Cho J., Garcya-Molina H., Page L.** Efficient crawling through URL ordering // Search and Indexing Techniques II: Proc. / 7th World-Wide Web Conf., Brishbone, Australia, April, 1998 (http://www7.scu.edu.au/programme/fullpapers/1919/com1919.htm).
4. **Davison B. D.** Topical locality in the Web // Research and Development in Information Retrieval: Proc. / SIGIR'00: 23rd Annual Internat. ACM SIGIR Conf., Athens, Greece, July 2000. — ACM, 2000. — P. 272–279 (http://citeseer.nj.nec.com/271585.html).
5. **Diligenti M., Coetzee F, Lawrence S., and other.** Focused crawling using context graphs //Very Large Data Bases: Proc. / VLDB 2000: 26th Internat. Conf., Cairo, Egypt, September 2000 — Morgan Kaufmann, 2000. — P 527–534 (http://www.informatik.uni-trier.de/~ley/db/conf/vldb/DiligentiCLGG00.html).
6. **Flake G., Lawrence S., Giles C. L.** Efficient Identification of Web Communities // Knowledge Discovery and Data Mining: Proc. /SIGKDD'00: 6th ACM Int'l Conf., Boston, MA, August 2000. — ACM, 2000. — P. 150–160.
7. **Kleinberg J., Gibson D., Raghavan P.** Inferring Web communities from link topology // Hypertext and Hypermedia: Proc. / 9th ACM Conf., Pittsburgh, Pennsylvania, USA, June, 1998. — ACM, 1998. — P 225–234 (http://citeseer.nj.nec.com/36254.html).
8. **Kleinberg J.** Authoritative sources in a hyperlinked environment // Discrete Algorithms : Proc. / ACM-SIAM Symp., San Francisco, California, USA, January, 1998. — ACM 1998. — P. 668–677.
9. **Лоуренс С.** Контекст при поиске в Интернет // Открытые системы. — 2000. — N 12. — (http://www.osp.ru/os/2000/12/062.htm).
10. **Хензингер М.** Анализ гиперссылок в Web // Открытые системы. — 2001. — N 10. — (http://www.osp.ru/os/2001/10/050.htm).
11. **Романова Е.В., Романов М.В., Некрестьянов И.С**. Использование интеллектуальных сетевых роботов для построения тематических коллекций // Тр. 1-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». — Санкт-Петербург, 1999 (http://www.dl99.nw.ru/PDF/16.pdf).