Исследование эффективности применения методов совместного анализа текстов и гиперссылок для поиска тематических сообществ

Козлов Д.Д., Белова А.А.

Факультет Вычислительной математики и кибернетики
МГУ им. М.В. Ломоносова
ddk@cs.msu.su, nastenka@lvk.cs.msu.su

Аннотация

Одним из важных аспектов тематического поиска в Web является создание у пользователя представления о том, какая имеется информация по интересующей его теме. Для этой цели могут применяться методы поиска тематических сообществ [8,9], основанные на анализе структуры гиперссылок. В данной работе исследуется эффективность методов поиска тематических сообществ, основанных на анализе гиперссылок (HITS, SALSA), а также комбинированных методов, сочетающих анализ гиперссылок с традиционными методами анализа текстов (TFIDF и LSA). Экспериментальные результаты показывают, что использование анализа гиперссылок стало менее эффективно из-за изменения структуры гиперссылок, а для эффективного поиска тематических сообществ требуется большое количество эвристик. Дополнительное применение эвристик и элементов анализа текста позволяет повысить качество работы методов поиска тематических сообществ. При этом применение методов, основанных на разложении по собственным векторам, не дает ощутимого выигрыша в качестве, а существенно уступает в вычислительной сложности.

1. Введение

Одним из частных случаев информационного поиска является так называемая задача тематического поиска [1], для которой характерно то, что

1) в начале поиска пользователь не знает четко свою информационную потребность, а имеет о ней лишь общее представление – тему. Поэтому он не может сформулировать запрос к информационно-поисковой системе, в ответ на который, будут выданы интересующие его объекты;
2) в процессе поиска пользователь уточняет свою информационную потребность. Результатом поиска является не только отбор нужных пользователю объектов, но и уяснение им самим своей информационной потребности.

Таким образом, важным аспектом тематического поиска является не только отбор нужных пользователю документов, но и обучение пользователя с целью уточнения им самим своей информационной потребности. Обучение пользователя и создание у него общей картины того, как представлена в Web интересующая его тема, особенно важны в начале поиска.

В настоящее время на практике задача тематического поиска в Web решается в основном с помощью последовательности взаимодействий пользователя с информационно-поисковой системой, чаще всего системой поиска по ключевым словам (search engine в английской терминологии, например, Яндекс или Google). При этом, с точки зрения системы поиска по ключевым словам, каждый запрос не связан с предыдущим, а с точки зрения пользователя, все запросы образуют единую логическую цепочку. В работе Бейтс [1] был сформулирован тезис, что каждый запрос к поисковой системе в последовательности взаимодействий имеет свое назначение и отражает изменяющуюся в процессе поиска информационную потребность. По Бейтс, пользователь сначала старается создать общую картину имеющихся ресурсов по данной теме (т.е. обучается), а потом начинает поиск с целью отбора наиболее интересной информации.

Для решения задачи обучения, т.е. создания общей картины имеющейся по данной теме информации могут, применяться так называемые методы поиска тематических сообществ [8,9], обнаруживающие тематические сообщества - группы web-страниц, связанных гиперссылками и относящихся к одной теме.

Задача поиска тематических сообществ

Понятие тематического сообщества не имеет общепринятого формального математического определения, поэтому в различных работах можно найти много определений тематического сообщества [8,9,14]. В большинстве работ тематическое сообщество Web-страниц определяется на основе анализа структуры гиперссылок между этими страницами. Методы поиска тематических сообществ обычно находят так называемое ядро сообщества – группу Web-страниц, в которой страницы много ссылаются друг на друга. При этом, ввиду отсутствия формального определения тематического сообщества, граница сообщества не определяется. Исключением является работа [8], в которой предложено формальное определение: тематическим сообществом называется множество страниц, в котором каждая страница имеет больше ссылок на другие страницы сообщества, чем на страницы вне сообщества.

Методы поиска тематических сообществ можно разделить на два класса:

- Методы, основанные на поиске структурного шаблона (сигнатуры сообщества) – подграфа графа Web определенного вида. Так, например, сигнатурой сообщества может быть NK-клан граф [14] или полный двудольный подграф [11].
- Методы, основанные на анализе графа гиперссылок средствами линейной алгебры [10] или статистическими средствами [2,12].

В работе Дэвисона [4] было экспериментально обосновано так называемое свойство тематической локальности Web, согласно которому вероятность того, что страницы, связанные гиперссылкой, относятся к одной теме выше, чем вероятность того, что две произвольные страницы, не связанные гиперссылкой, относятся к одной теме. На основе свойства тематической локальности можно предположить, что страницы внутри такой группы относятся к одной теме, т.е. образуют тематическое сообщество.

Большинство методов поиска тематических сообществ в Web используют следующие предположения о семантике гиперссылок:

1) Когда автор ставит в своем документе A ссылку на чужой документ B, он рекомендует читателю A прочитать еще и B.

2) Если два документа соединены ссылкой, то вероятность того, что они относятся к одной теме выше, чем в случае отсутствия ссылки.

Такое предположение было весьма правдоподобно в то время, когда в Web не было рекламы, сейчас же распространение баннерных сетей, счетчиков и автоматически создаваемых ссылок существенно снижает количество страниц, соответствующих этому предположению, и существенно понижает качество работы этих методов [5].

В связи с этим, одним из направлений развития методов поиска тематических сообществ является совмещение анализа структуры гиперссылок с анализом текстов страниц. Так, в работе [3] было показано, что использование анализа текста лишь некоторых элементов страниц позволяет существенно повысить качество поиска.

В данной работе сделана попытка экспериментального анализа эффективности методов поиска тематических сообществ, основанных на совместном анализе текстов страниц и структуры гиперссылок, и их сравнение с методами, основанными на анализе структуры гиперссылок.

В данной работе также описана модель тематического поиска, основанная на идее методов поиска тематических сообществ и существенной интерактивности тематического поиска по Бейтс.

2. Идея исследования

Наиболее широко распространенным методом поиска тематических сообществ является алгоритм HITS [10]. Этот метод осуществляет анализ структуры гиперссылок подграфа графа Web на основе разложения матрицы смежности по собственным векторам. В работе Клейнберга и последующих работах акцент, в основном, сделан на использовании главного собственного вектора. В то же время самим Клейнбергом в [10] была указана возможность использования неглавных собственных векторов.

С другой стороны, в работе Дэвисона [6] показана похожесть метода HITS и латентно-семантического анализа [7], что породило идею унифицированного метода поиска тематических сообществ, совмещающего анализ структуры гиперссылок и анализ текста.

Как было указано в предыдущем разделе, ряд исследований показывает, что совмещение анализа структуры гиперссылок и анализа текста позволяет повысить точность работы методов поиска тематических сообществ и является перспективным направлением развития методов поиска тематических сообществ.

В данной работе производится сравнение эффективности четырех методов поиска тематических сообществ:

- на основе классического алгоритма HITS, как исходной точки отсчета при сравнении,
- на основе HITS с использованием неглавных собственных векторов,
- на основе комбинирования HITS и латентно-семантического анализа [10],
- на основе комбинирования [15] анализа гиперссылок с помощью SALSA [12] и анализа текста с помощью TFIDF [13].

Сравнение эффективности методов поиска тематических сообществ в данной работе производится на основе реальных данных, взятых из русскоязычной части Web. Такое сравнение позволит также оценить степень применимости методов поиска тематических сообществ к существующей структуре гиперссылок в русскоязычной части Web. Этот вопрос является открытым в связи с тем, что методы поиска тематических сообществ разрабатывались исходя из описанных выше предположений о семантике гиперссылок, а широкое внедрение в поисковые системы алгоритма PageRank привело к обратному результату: Web стал изобиловать

автоматически генерируемыми ссылками, которые искусственно поднимают рейтинг некоторых сайтов в поисковых системах.

3. Описание методов и экспериментов

В данном разделе описаны исследуемые в данной работе методы поиска тематических сообществ: HITS (базовый метод, с которым традиционно сравниваются все остальные), HITS с использованием неглавных собственных векторов и два метода, совмещающие анализ структуры гиперссылок и анализ текста: метод, основанный на идее Дэвисона [6] о совмещении HITS и LSA, метод [15], основанный на совмещении SALSA и TFIDF. Описана постановка экспериментов основанная на идеях TREC WebTrack, приведены результаты экспериментов.

Алгоритм HITS

В основе алгоритма HITS лежит понятие значимости страницы. Наиболее значимыми страницами предложено считать те страницы, на которые больше всего ссылаются другие значимые страницы. Такие страницы называются первоисточниками (authorities). Первоисточники являются наиболее значимыми в рамках заданной темы, поэтому на них часто ссылаются другие страницы, относящиеся к данной теме. Это свойство позволяет выявить так называемые индексные страницы (hub pages), которые ссылаются на несколько первоисточников, относящихся к одной теме. Вместе оба типа значимых страниц образуют отношение взаимного усиления (mutually reinforcing relationship), то есть качественный первоисточник ссылается на много качественных индексных страниц и качественная индексная страница ссылается на много качественных первоисточников. Таким образом, целью анализа в HITS является поиск наиболее качественных первоисточников и наиболее качественных индексных страниц.

Работа алгоритма строится в два этапа. На первом этапе строится так называемый сфокусированный подграф Web, который содержит страницы, полученные путем посылки запроса системе поиска по ключевым словам (СПКС). На втором этапе производится анализ сфокусированного подграфа Web и вычисляются наиболее значимые документы.

Набор связанных гиперссылками страниц можно рассматривать как ориентированный граф $G=(V,E)$. На первом этапе по заданному пользователем запросу σ вычисляется подграф Web, с которым

будет работать вторая часть алгоритма. Пусть S_σ – множество вершин этого подграфа. К S_σ предъявляются следующие требования:

- относительно маленький размер (тысячи страниц);
- наличие большого количества релевантных запросу страниц;
- наличие большинства наиболее значимых страниц (авторитетных и индексных).

Множество S_σ строится следующим образом. Вначале запрос σ посылается системе поиска по ключевым словам и берутся первые t результатов. Они образуют базовое множество R_σ . Затем S_σ полагается равным R_σ, далее к S_σ добавляются все страницы, на которые ссылаются страницы из R_σ, и все документы, которые ссылаются на страницы из R_σ.

В ходе построения $G[S_\sigma]$ – подграфа графа Web, индуцированного на множестве S_σ, для повышения качества работы алгоритма применяется ряд эвристик, позволяющих отличить семантически значимые ссылки.

1) Наличие большого количества гиперссылок на страницу свидетельствует о популярности, а не о релевантности. Следовательно, если очень много страниц ссылаются на некоторую страницу p, то все гиперссылки включать не надо, чтобы сохранить малый размер S_σ.
2) На основании принадлежности к домену все гиперссылки между страницами в $G[S_\sigma]$ разделяются на 2 класса: внутренние и внешние. Внутренние гиперссылки соединяют страницы внутри одного домена. Внешние гиперссылки соединяют страницы из разных доменов. Все внутренние гиперссылки удаляются.
3) Страница, на которую ссылаются много страниц с одного сайта (или из одного домена), скорее всего не содержит релевантной информации, а является рекламной, навигационной и т.п. Для таких страниц число ссылок с данного сайта (из домена) ограничивается (m = 4-8). Страницы, для которых этот порог превышен, удаляются.

На втором этапе происходит анализ графа $G[S_\sigma]$. Каждой странице сопоставляются два веса: x – *AP-вес*, показывающий качество страницы как первоисточника, и y – *HP-вес*, показывающий

качество страницы как индексной страницы. Весам дается произвольное ненулевое начальное приближение и затем производится итерационный процесс, состоящий из последовательного применения двух операций I и O.

Операция I вычисления AP-веса для страницы p:

$$x\langle p\rangle \leftarrow \Sigma_{q:\,(q,p)\in E}\ y\langle q\rangle.$$

Операция O вычисления HP-веса для страницы p:

$$y\langle p\rangle \leftarrow \Sigma_{q:\,(p,q)\in E}\ x\langle q\rangle.$$

Обоснование сходимости итерационного процесса приведено в [5]**.**, где также указано, для того чтобы 20 наибольших значений AP-весов и HP-весов становились стабильными достаточно около 20 итераций.

В качестве ядра тематического сообщества берутся наиболее значимые первоисточники.

Использование неглавных собственных векторов в HITS

Описанный выше итерационный процесс имеет матричную форму записи: $x = M^T y$, $y = Mx$, где M – матрица смежности графа G, откуда видно, что вектора x и y являются главными собственными векторами матриц M^TM и MM^T. В [10] также отмечена возможность использования неглавных собственных векторов для определения дополнительных сообществ, например, при неоднозначности запроса. При этом каждый неглавный собственный вектор соответствуют двум кластерам: один кластер – наибольшие положительные значения, а другой кластер – наименьшие отрицательные. В экспериментах при подсчете результатов брались 10 наибольших и 10 наименьших значений из первых трех собственных векторов.

Метод унифицированного анализа гиперссылок и текста

Как было показано в предыдущем разделе, анализ графа в HITS в матричном виде сводится к поиску собственных векторов матриц AA^T и A^TA. Обе эти матрицы имеют одинаковый набор собственных значений, и имеется связь между сингулярным разложением матрицы A и разложениями матриц AA^T и A^TA по собственным векторам. Если $A=VDU^T$ – сингулярное разложение матрицы A, то $A^TA=UD^TV^T\ VDU^T= UD^TV^T\ VDU^T=U\ (D^TD)U^T$, что, учитывая диагональную структуру D, есть разложение по

собственным векторам матрицы A^TA, аналогично для AA^T.
Метод анализа текстов на основе латентно-семантического анализа [7] также использует аппарат сингулярного разложения, только в отличие от матрицы смежности графа используется матрица термы на документы, элементами которой являются частотные характеристики вхождения термов в документы.

В работе Дэвисона [6] было предложено использовать для совмещения анализа текстов и анализа структуры гиперссылок блочную матрицу специального вида:

$$M = \begin{matrix} \text{term x term} & \text{term x doc} \\ \text{doc x term} & \text{doc x doc} \end{matrix}$$

где doc x term - матрица частотных характеристик вхождения термов в документы (Web-страницы), term x doc - транспонированный вариант предыдущей матрицы, матрица term x term - отражает зависимости между термами (для независимых термов она нулевая), а матрица doc x doc - отражает гиперссылки между документами, аналогично алгоритму HITS.

Сингулярное разложение блочной матрицы позволяет выделить кластеры, которые в данном случае состоят не только из документов, но и из термов. Таким образом, можно получить именованные термами кластеры документов. Однако до осени 2004 года не было опубликовано исследований этого метода. Потенциальное преимущество применения этого метода для поиска тематических сообществ состоит с том, что с его помощью можно не только выделить тематические сообщества, но и выделить характеризующие эти сообщества термы.

Метод совместного анализа текста и структуры гиперссылок

Следующий метод [15] основан на совмещении анализа структуры гиперссылок с помощью алгоритма SALSA [12] и средств анализа текста на основе алгоритма TFIDF [13].

Первый этап работы метода аналогичен алгоритму HITS: исходная формулировка темы посылается системе поиска по ключевым словам, которая возвращает базовое множество страниц, а на его основе строится подграф графа Web. При расширении базового множества используется анализ текста исходящих гиперссылок: в

расширенное множество добавляются не только страницы с других сайтов, но и страницы типа «карта сайта», «ссылки», «заглавная страница». Количество страниц добавляемых с одного сайта ограничено, а приоритет имеют страницы, в ссылках на которые встречаются термины запроса. Здесь же используются две эвристики для исключения баннерной рекламы: первая эвристика проверяет ссылку на принадлежность к баннерной сети по «черному» списку, построенному на основе каталога Яндекса, а вторая отсекает баннеры по шаблону <img> … .

После построения расширенного множества и построения G – подграфа графа Web индуцированного на расширенном множестве производится укрупнение элементов подграфа: страницы объединяются в группы, называемые ресурсами. Ресурсами могут быть сайт, часть сайта, домашняя страница пользователя, группа сайтов. Ресурс должен быть тематически однороден. Для каждого ресурса определяется заглавная страница, которая является точкой входа в ресурс. Например, для пользователя ищущего информацию о программировании на java, заглавной страницей будет www.sun.ru/java, а не www.sun.ru, так как только часть сайта www.sun.ru посвящена программированию на java.

Объединение страниц в ресурсы производится по следующим правилам:

1) Изначально каждому сайту сопоставляется свой ресурс.
2) Каждая домашняя страница пользователя (www.example.org/~username) выделяется в отдельный ресурс. Все страницы, составляющие домашнюю страницу пользователя (например, www.example.org/~username/index.html, www.example.org/~username/links.html и т.д.), относятся к одному ресурсу.
3) Сайты, имеющие общий домен не менее чем второго уровня, объединяются в один ресурс в том случае, если внутри такой группы каждый сайт ссылается на всех остальных членов группы. Это правило позволяет отсеивать такие группы как, например, www.juga.ru, lib.juga.ru, list.juga.ru и т.д., так как подобные методы организации сайтов направлены на препятствие корректному анализу гиперссылок.
4) Если одна из страниц сайта имеет больше входящих внешних ссылок, чем корневая страница сайта, то сайт разбивается на два ресурса.

Заглавная страница ресурса определяется по следующим правилам:

1) Для ресурса, представляющего собой домашнюю страницу пользователя, например, www.example.org/~username, заглавной страницей является сама домашняя страница.
2) Для ресурса, представляющего собой целый web-сайт, заглавной страницей будет являться заглавная страница сайта.
3) Для ресурса, не содержащего заглавную страницу сайта, заглавной является страница с наименьшей глубиной пути в URL.
4) Для ресурса, представляющего собой группу сайтов с общим именем домена domain.ru, заглавной является заглавная страница сайта www.domain.ru, если таковой есть, а в противном случае – заглавная страница сайта, содержащего наибольшее число входящих внешних ссылок.

После построения множества ресурсов на основании графа G строится граф ресурсов. Ребра, соединяющие в графе G страницы, относящиеся к разным ресурсам, отображаются в ребра графа ресурсов. Использование графа ресурсов вместо графа страниц позволяет сократить размер анализируемого графа, а также сократить в графе количество ребер, соответствующих гиперссылкам, не удовлетворяющим предположениям о семантике.

Для вычисления оценок значимости предлагается алгоритм, объединяющий анализ структуры гиперссылок на основе алгоритма SALSA и анализ текста. Алгоритм SALSA для связного графа совпадает с первым шагом HITS (обоснование см. в [2]), но при этом SALSA в отличие от HITS не обладает столь выраженным эффектом смещения темы. Вычисление оценок значимости в SALSA производится следующим образом. Пусть N_j – количество вершин графа в j-ой связной компоненте графа. N – количество вершин графа. A_i – АР-вес ресурса i, принадлежащего j-ой связной компоненте. B_j – сумма всех весов входящих ребер по j-ой компоненте. B(i) – сумма весов всех ребер, входящих в вершину i. Тогда $A_i = N_j * B(i) / (N * B_j)$. Аналогично вычисляются HP-веса: $H_i = N_j * F(i) / (N * F_j)$. Вес ребра вычисляется путем сложения следующих составляющих:

1) Const0 (=1);
2) Const1 при наличии ресурса, из которого выходит ребро, в списке, полученном от системы поиска по ключевым словам;
3) Const2 * релевантность текста ссылки, соответствующей ребру

(только при вычислении AP-веса (B));
4) Const3 * средняя релевантность страниц ресурса, на который направлено ребро (только при вычислении HP-веса (F)).

Для анализа текста используется алгоритм TFIDF [13]. В качестве TFIDF-модели выступает строка запроса, при этом предполагается, что в дальнейшем в процессе интерактивного поиска (см. ниже раздел про модель интерактивного тематического поиска) в качестве TFIDF-модели будут использоваться страницы, указываемые пользователем.

Формирование тематического сообщества производится на основе взвешенной суммы трех весов: AP-веса, HP-веса и оценки релевантности. В тематическое сообщество отбирается 10 ресурсов с наибольшей суммарной оценкой.

Постановка экспериментального исследования

Целью экспериментального исследования являлся анализ эффективности рассматриваемых методов поиска тематических сообществ, а именно:
1) сравнение точности результатов поиска тематических сообществ;
2) сравнение вычислительной сложности методов.

Следуя рекомендациям TREC WebTrack по оценке методов тематического ИП в Интернет [16], в данной работе экспериментальное исследование проводилось на задаче поиска ключевых ресурсов по заданной теме (в английской терминологии Topic Distillation). Задача поиска ключевых ресурсов заключается в том, что по заданному в виде набора ключевых слов определению темы необходимо найти 10 ключевых ресурсов Web, относящихся к этой теме. Содержательно, ключевой ресурс – это web-сайт, который является источником документов, относящихся к данной теме. «Ключевой» здесь означает наиболее ценный в смысле количества и качества документов. Формально, ключевой ресурс может быть сайтом или частью сайта, относящейся к заданной теме, или отдельной страницей, содержащей ссылки на ресурсы по данной теме. Ключевые ресурсы не должны быть вложенными друг в друга.

Традиционно экспериментальные исследования ИПС проводятся на основе методов экспертной оценки на заранее подготовленном

наборе данных. Мерами оценки эффективности обычно являются точность и полнота поиска. Однако в настоящее время тестовые наборы русскоязычных данных для этой задачи отсутствуют, а их создание чрезвычайно трудоемко. Это приводит к необходимости проведения исследования не на модельных данных, а на реальных данных русскоязычной части Web. Использование реальных данных не позволяет корректно оценивать полноту поиска ввиду большого объема и динамичности Web. В рамках данной работы в качестве меры эффективности использовалась точность поиска. Инструментом измерения являлся метод экспертной оценки. Наиболее качественными экспертными оценками в русскоязычной части Web являются составленные вручную каталоги, например, list.ru, yaca.yandex.ru, aport.ru.

В процессе подготовки экспериментальных данных темы[1] выбирались из листьев каталога yaca.yandex.ru. Для построения экспертных оценок использовались объединенные в единый список ключевых ресурсов данные из популярных российских каталогов list.ru, yaca.yandex.ru, aport.ru. Размер списка ключевых ресурсов почти всегда превышал 10 ресурсов.

Оценка результатов проводилась следующим образом. Для каждой информационной потребности осуществлялся поиск с помощью исследуемых алгоритмов. Первые 10 результатов поиска сравнивались с экспертными оценками. Если страница или сайт, выданные системой, попадали в множество релевантных ресурсов, указанное экспертом, то результат считался релевантным, в противном случае – нет. Оценкой точности поиска являлось количество релевантных результатов среди первых 10 результатов поиска.

Результаты экспериментов

Для экспериментов были использованы следующие темы

- ландшафтный дизайн (q1);
- авторская песня (q2);
- программирование на java (q3);
- технология xml (q4);
- операционная система linux (q5);

[1] Формулировки тем отличаются от формулировок в каталогах, так как в каждом из каталогов своя классификация.

- раннее развитие детей (q6);
- обзоры железа (q7);
- пчеловодство (q8).

Диаграмма 1. Сравнение точности базового варианта HITS и HITS с укрупнением анализируемых объектов.

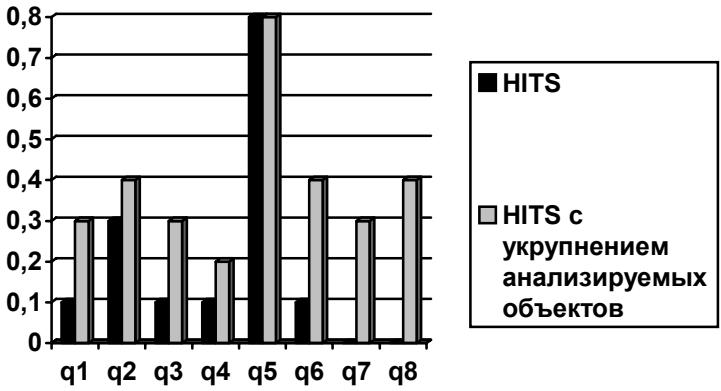


Диаграмма 2. Сравнение базового варианта HITS и HITS с использованием неглавных собственных векторов.

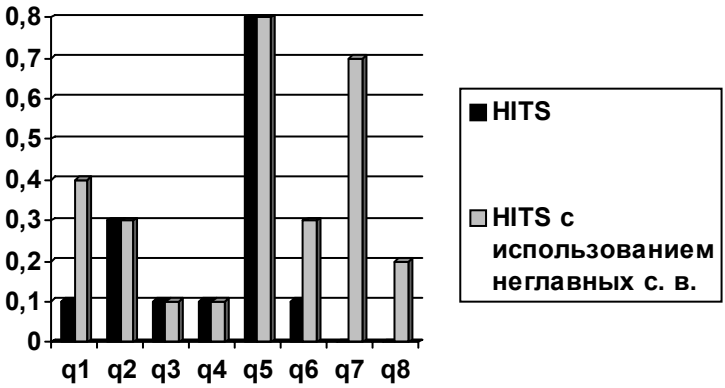
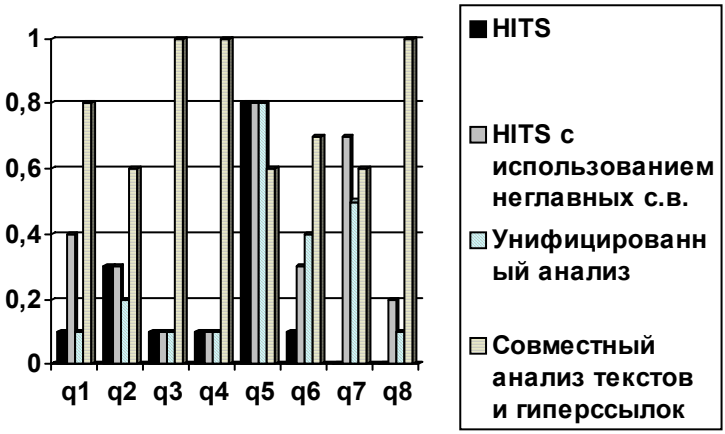


Диаграмма 3. Сравнение точности всех четырех алгоритмов.



В большинстве случаев алгоритм HITS показал низкое качество результатов поиска в силу своей неустойчивости к зашумленным входным данным: почти для каждой темы в начальном графе присутствовали тематические сообщества, не относящиеся к теме.

Если в графе G присутствовало только одно сообщество, которое соответствовало теме, то HITS давал очень хорошие результаты, примером тому является тема «операционная система linux». Это согласуется с результатами исследований [2]. В то же время анализ текста, используемый в последнем методе, позволял отфильтровать сообщества, которые не имеют отношения к теме.

Применимость алгоритма HITS существенно ограничивает наличие внутридоменных навигационных ссылок, например, lib.juga.ru, forum.juga.ru, www.juga.ru, между которыми существует полный граф ссылок. На таких структурах HITS сразу останавливается, в то время как использование ресурсов в последнем методе позволяет обойти эту проблему.

Для всех алгоритмов существенную проблему создавали также баннеры: без отсечения ссылок, соответствующих баннерам, алгоритмы не работали вовсе, поэтому при всех сравнениях алгоритмов использовались эвристики, описанные выше.

Особенность русскоязычной части Web состоит в том, что для многих тематических сообществ наиболее авторитетными страницами являются англоязычные. По многим темам, особенно по информационным технологиям, страницы содержат намного меньше ссылок на русскоязычные сайты, чем на англоязычные сайты. Можно предположить, что для любого другого национального языка ситуация аналогична. Это порождает существенную проблему при анализе: если в граф G попадают англоязычные страницы, то они доминируют, а если они удаляются, то происходит потеря важной информации о связях русскоязычных страниц, совместно цитирующих англоязычную страницу. В данной работе англоязычные страницы удаляются из графа G, в результате чего возможна потеря качества оценок значимости.

В некоторых экспериментах результаты HITS с неглавными собственными векторами существенно превосходят результаты HITS. Это происходит, когда главный собственный вектор зацикливается на несемантических ссылках, например, на нескольких внутренних доменах *.rambler.ru. Эта ситуация отчасти исправляется эвристиками по укрупнению анализируемого графа.

В большинстве случаев наиболее хорошо показал себя метод с применением анализа текста. При этом использование

унифицированного анализа оказалось не оправдано. Особенно сложно оказалось добиться сочетания вкладов различных блоков матрицы. В большинстве экспериментов получалось доминирование матрицы ссылок.

Показательно, что с использованием эвристик от HITS можно было добиться весьма неплохого результата, сопоставимого с методами, с применением анализа текста. Это обусловлено малым размером TFIDF-модели. Для эффективного применения TFIDF необходимо обучение алгоритма, например, на основе обратной связи с пользователем.

Оценка вычислительной сложности методов

Для сравнения оценки вычислительной сложности рассматриваемых методов использовался следующий подход. Все системы тематического поиска, основанные на анализе подграфа графа Web, рассматриваются как агенты. Основные операции, производимыми такими агентами, можно классифицировать следующим образом:

- получение заданной страницы из Интернет и получение ссылок с нее;
- получение от системы поиска по ключевым словам ссылок на заданную страницу;
- получение от системы поиска по ключевым словам списка страниц по запросу q;
- вычисление оценки релевантности для страницы;
- вычисление оценки релевантности для текста гиперссылки;
- вычисление оценок значимости;
- остальные операции.

Использование такой классификации позволяет сравнивать эффективность целого класса систем тематического поиска, таких как [3,8,9,10,12,15] за счет рассмотрения их как агентов, которые в процессе своей работы осуществляют операции описанных классов. Вычислительная сложность алгоритма при этом вычисляется как количество операций разных классов. Для класса "остальные операции" необходимо отдельно указывать оценки сложности. Чтобы перевести набор из нескольких оценок в единую оценку, нужно указать среднюю вычислительную сложность (или временную оценку) для каждого класса операций.

При использовании внешней системы поиска по ключевым словам для первых трех классов операций основной временной затратой будет скачивание результата с удаленного сервера. Для него можно вычислить среднее значение времени скачивания, основываясь на среднем размере скачиваемой страницы.

Время вычисления оценок релевантности также зависит от среднего размера анализируемой страницы и используемого алгоритма. Зная оценку сложности для алгоритма и средний размер страницы, можно получить временную оценку. Оценка релевантности текста гиперссылки аналогична, но средняя длина текста ссылки существенно меньше, чем длина текста. Сложность оценки значимости зависит от количества вершин анализируемого графа. Таким образом, для приближенной оценки вычислительной сложности необходимы следующие настроечные параметры: средний размер страницы, средняя длина гиперссылки, среднее время скачивания страницы, размер анализируемого графа.

На практике при использовании внешней системы поиска по ключевым словам основное время в процессе работы систем тематического поиска занимает получение информации по сети, а не локальная обработка информации. Из локальных действий наиболее длительным является анализ текста.

Сравнение вычислительной сложности рассматриваемых методов приведено в таблице 1:

Таблица 1. Сравнение вычислительной сложности методов

Алгоритм / Класс операций	HITS	HITS EV	Униф. анализ	Комбинированный анализ
Получение от СПКС результатов по запросу q	1	1	1	1
Скачивание страниц	N	N	N	N
Скачивание ссылок на страницу	|R|	|R|	|R|	|R|
Вычисление оценки релевантности страницы	0	0	0	0
Вычисление оценки	0	0	0	~ 7 * N

релевантности текста гиперссылки				
Вычисление оценки значимости	1 раз, $O(N^2)$ сложений и умножений	1 раз, $O(N^3)$ сложений и умножений	1 раз, $O((N+$среднее число термов в документе$*N)^3)$ сложений и умножений	1 раз, $O(N^2)$ сложений и умножений
Удаление внутренних гиперссылок	O(N)	O(N)	O(N)	0
Построение графа ресурсов	0	0	0	O(N)

N – размер анализируемого графа (как правило 1000-5000).
|R| - размер базового множества (50-200).

Сравнение вычислительной сложности производилось в предположении, что разложение по собственным векторам строится полностью.

Модель интерактивного тематического поиска

На основе идеи Бейтс об интерактивной природе тематического поиска и на базе методов поиска тематических сообществ можно предложить следующую модель интерактивного тематического поиска.

Пусть имеется пространство поиска:

$G = <O, L>$ – ориентированный граф объектов,
где O – множество объектов поиска, а L – множество ориентированных ссылок между объектами.
Каждый объект $o_i \in O$ обладает вектором атрибутов:

$$O = \{o_j = <a_{j1},\ldots, a_{jk}>, k=k(j)\},$$

Для разных объектов наборы атрибутов могут различаться.
Пространство поиска в каждый момент времени представляет собой лишь некоторое приближение всего графа Web. Это приближение может изменяться в процессе поиска:

$$G_t = <O_t, L_t> = G(t), t=0,1,2,\ldots$$

Пользователь формулирует свое начальное представление об информационной потребности в виде начального запроса:

$$b_0=\{b_{01}, \ldots, b_{0m}\},$$

который задает множество требуемых атрибутов объектов поиска. Запрос пользователя изменяется в процессе поиска $b_t = b(t)$.

Функция поиска A осуществляет информационный поиск в пространстве G_t по запросу b_t:

$$A(b_t, G_t) = \bar{o} \subseteq O_t.$$

Результатом поиска является множество $\bar{o}$ объектов поиска из пространства поиска O_t.

Пользователь оценивает результаты поиска на соответствие своей информационной потребности:

оценка результатов поиска $E(b_t, \bar{o}) = \bar{o}' \subseteq \bar{o}$.

В процессе поиска пользователь уточняет свою информационную потребность, и на основе уточненной информационной потребности строится уточненный запрос и выбирается направление поиска:

функция уточнения запроса $Q(\bar{o}', b_t) = <b_{t+1}, d_{t+1}>$. $d_{t+1} \subseteq O_t$.

Новое направление поиска задается в виде множества страниц, ссылки на которые (и с которых) надо раскрыть при расширении пространства поиска. На основании выбранного направления поиска производится расширение пространства поиска:

функция расширения пространства поиска: $G_{t+1} = W(G_t, d_{t+1}, b_t)$.

Применительно к поиску в Web объектами поиска являются web-страницы, атрибутами которых являются слова, встречающиеся в текстах страниц. Для расширения пространства поиска используются следующие операции:

1) получение страницы, заданной URL;
2) получение ссылок на страницу, заданную URL, путем обращения к системе поиска по ключевым словам;
3) получение от системы поиска по ключевым словам списка страниц по запросу из набора ключевых слов.

Эта модель обобщает методы поиска тематических сообществ для случая интерактивного поиска с обратной связью.

4. Выводы и обсуждение результатов

1. Эксперименты показали, что в настоящее время методы поиска тематических сообществ, основанные на указанных выше предположениях о семантике гиперссылок становятся неприменимы без использования дополнительных эвристик, отсеивающих гиперссылки, заведомо не удовлетворяющие предположениям о семантике. Для решения этой задачи может применяться укрупнение анализируемых объектов, например, использование ресурсов, определенных выше, вместо страниц. Этот подход позволяет снизить количество попаданий в анализируемый граф гиперссылок, не удовлетворяющих предположениям о семантике, а также позволяет незначительно снизить размер анализируемого графа (для HITS это не изменяет линейности сложности).

По результатам экспериментов можно сделать вывод, что на реальных данных русскоязычной части Web алгоритмы поиска тематических сообществ показали себя заметно хуже, чем в экспериментах [4-6].

2. Проведенные эксперименты показали, что сочетание анализа текста с анализом структуры гиперссылок повышает точность поиска, но ценой увеличения сложности. В повышении точности также играет роль использование ресурсов вместо страниц при анализе.

3. По результатам проведенных экспериментов можно сделать вывод, что применение унифицированного метода анализа текста и структуры гиперссылок является вычислительно сложным и при этом не дает существенного выигрыша в качестве поиска.

4. Проведенные эксперименты показали, что использование неглавных собственных векторов в HITS при существенном повышении вычислительной сложности (необходимо строить разложение матрицы большого размера) по сравнению с HITS не дает большого выигрыша в точности. Отчасти этот результат может быть связан со структурой исходных данных – в используемых экспериментальных данных в одном анализируемом графе не содержится нескольких тематических сообществ. Здесь возможны несколько путей развития. Во-первых, можно пробовать снижать вычислительную сложность разложения путем построения не полного разложения матрицы смежности, а выделения только 2-3

наибольших собственных значений. Во вторых, при наличии в графе нескольких тематических сообществ, возможно последовательно комбинировать SALSA и HITS. При этом SALSA будет выделять главные страницы из разных сообществ, а затем с помощью HITS, изменяя веса в матрице, можно будет выделить каждое из сообществ в отдельности. Этот подход не выходит за пределы линейной сложности.

5. Постоянно изменяющаяся структура гиперссылок в Web (оптимизация сайтов, реклама и т.д.) заставляет использовать все новые и новые эвристики для обеспечения применимости методов поиска тематических сообществ к Web. Методы поиска тематических сообществ более эффективно использовать на незашумленных данных, например на графе библиографического цитирования научных статей (например, на графе CiteSeer).

5. Литература

[1] Bates M., The design of browsing and berrypicking techniques for the online search interface. Online Review 13, 5, 1989

[2] Borodin A., Roberts G., Rosenthal J., Tsaparas P. Finding Authorities and Hubs From Link Structures on the World Wide Web. Tenth World Wide Web Conference, Hong Kong, 2001

[3] Bharat K., Henzinger M. Improved Algorithms for Topic Distillation in a Hyperlinked Environment. ACM SIGIR conference on Research and Development in IR, 1998.

[4] Davison B. Topical locality in the Web. Proceedings of the ACM SIGIR'2000 Conference, 2000.

[5] Davison B., Recognizing Nepotistic Links on the Web, AAAI Workshop on Artificial Intelligence for Web Search, 2000.

[6] Davison B., Unifying Text and Link Analysis, IJCAI Workshop on Text-Mining & Link-Analysis, 2003.

[7] Deerwester S., et al. Indexing by Latent Semantic Analysis, Journal of the American Society of Information Science, 1990.

[8] Flake G., Lawrence S., Giles C., Coetzee F. Self-Organization of the Web and Identification of Communities. IEEE Computer, 35(3), pp 66-71, 2002.

[9] Gibson D., Kleinberg J., Raghavan P. Inferring Web communities from link topology. Proc. 9th ACM Conference on Hypertext and Hypermedia, 1998.

[10] Kleinberg J. Authoritative sources in a hyperlinked environment. Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998.

[11] Kumar S., Raghavan P., Rajagopalan S., Tomkins A. Trawling the Web for emerging cyber-communities. Eighth World Wide Web Conference, Toronto, Canada, May 1999.
[12] Lempel R., Moran S. The Stochastic Approach for Link-Structure analysis (SALSA) and the TKS Effect. Ninth World Wide Web Conference, 2000
[13] Salton G., Buckley C., Term Weighting Approaches in Automatic Text Retrieval, Cornell University Technical Report 87-881
[14] Terveen L., Hill W., Amento B. Constructing, Organizing, and Visualizing Collections of Topically Related Web Resourses. ACM Transactions on Computer-Human Interaction, Vol 6, No 1, March 1999, Pages 67-94
[15] Козлов Д.Д. Проблемы применения методов поиска тематических сообществ к задаче тематического информационного поиска в Интернет. Труды Всероссийской научной конференции Методы и средства обработки информации. – Москва: 2003, с. 211-215.
[16] TREC-2002 Web Track Guidelines, TREC, 2002.

Comparison of topic distillation methods based on links and text analysis

Kozlov D.D., Belova A.A.
Moscow State University Computer Science department

In this paper four approaches to topic distillation are compared: classical HITS [10], HITS with non-principal eigenvectors[9], unified text and link analysis [6] and combined analysis [15] based on SALSA, TFIDF and heuristics. Comparison is based on TREC WebTrack methodology but is made on real data from Russian part of the Web. The result is that topic distillation methods don't work without heuristics on modern Web, combination of text analysis and simple SALSA is better than complex unified analysis, HITS with enough heuristics is comparable with more complex methods with text analysis.