УДК 51:330.115

А. С. КУЦЕНКО, д-р. техн. наук,
К. А. СОКОЛИНСКИЙ

ВЫДЕЛЕНИЕ ПОИСКОВОГО СПАМА НА ОСНОВЕ МЕРЫ ССЫЛОЧНОЙ СХОЖЕСТИ ВЕБ-СТРАНИЦ

У статті описується міра посилальної схожості веб-сторінок і запропонований простий алгоритм для виділення кластерів веб-сторінок, підозрілих з точки зору використовування посилального спама. Кластеризація грунтується на зваженому графі схожості сторінок, який може бути одержаний з орієнтованого графа зв'язків веб-сторінок.

В статье описывается мера ссылочной схожести веб-страниц и предложен простой алгоритм для выделения кластеров веб-страниц, подозрительных с точки зрения использование ссылочного спама. Кластеризация основывается на взвешенном графе схожести страниц, который может быть получен из ориентированного графа связей веб-страниц.

Введение. Бурное развитие Интернета в последние годы привело к тому, что «всемирная паутина» стала незаменимым источником информации во многих отраслях человеческой деятельности. Для многих пользователей поисковые системы стали отправной точкой при поиске информации в сети. Соответственно, поисковые машины стали одним из основных источников посетителей для многих сайтов. Для коммерческих сайтов увеличение количества посетителей оборачивается ростом прибыли [1].

Большинство поисковых сервисов возвращают отсортированный по релевантности запросу список сайтов в качестве результатов. Количество сайтов в результате поиска может доходить до сотен тысяч и для многих запросов пользователи просматривают только первые 10 [1]. Это ведет к тому, что создатели некоторых сайтов пытаются обмануть поисковые машины и искусственно поднять рейтинг своих сайтов в результатах поиска. Данное явление получило название «поискового спама» [2].

Первые поисковые системы использовали разнообразные методы информационного поиска для нахождения релевантных запросу документов. Позже было обнаружено, что эти методы не достаточны для получения ценной информации из сети, т.к. они не учитывают связь между документами на основе гиперссылок. Большинство современных поисковых систем также используют алгоритмы, основанные на анализе гиперссылок, для улучшения ранжирования результатов поиска.

Наиболее известными алгоритмам, используемыми для ссылочного ранжирования, являются PageRank[3] и HITS[4], которые анализируют структуру гиперссылок и рассчитывают рейтинг страницы на основе входящих и исходящих ссылок. Было показано [2], что оба алгоритма являются уязвимыми к «ссылочному спаму». Данный вид поискового спама основан на создании искусственных веб-страниц, повышающих рейтинг

необходимой страницы. Разновидностью данного вида спама также является создание ссылочных сетей, используемых для повышения рейтинга всех либо некоторых страниц из сети. Подробный анализ некоторых видов ссылочного спама можно найти в [5].

Метод, предложенный в данной статье, основывается как на личных наблюдениях структуры страниц и сайтов которые могут считаться спамом, так и на анализе и предположениях, сделанных в нескольких источниках [1, 5, 6, 7]. Целью данного метода является выделение групп веб-страниц (веб-сайтов) произвольной конфигурации, которые могут быть подозрительны с точки зрения использования ссылочного спама. Мы используем слово «подозрительны», т.к. отнесение веб-страницы (веб-сайта) к категории спама является субъективной мерой, которая может разниться от одной поисковой машины к другой.

Постановка задачи

Введем представление множества страниц связанных гиперссылками в виде графа $G = (V, E)$, где V множество веб-страниц, а E множество ориентированных ребер $< i, j >$. Множество E содержит ребро $< i, j >$, если страница i ссылается на страницу j. Данному графу соответствует матрица смежности A_{ij}, где $i, j \in V$. Для данной матрицы $A_{ij} = 1$, если множество E содержит ребро $< i, j >$, и $A_{ij} = 0$ в противном случае.

На основе матрицы смежности графа G необходимо найти меру ссылочной схожести двух страниц и разработать алгоритм, позволяющий находить скопления (кластеры) схожих страниц.

Последующее изложение основано на некоторых предположениях и упрощениях. Будут рассматриваться только внешние гиперссылки, т.е. ссылки, ведущие за пределы доменного имени, в котором находится данная страница. Если страница содержит несколько идентичных ссылок, они будут рассматриваться как одна.

Мера ссылочной схожести

Использую матрицу смежности A_{ij}, мы можем получить множества исходящих $out(i)$ и входящих $in(i)$ ссылок для любой вершины $i \in V$. Множество исходящих ссылок содержит страницы из множества V, на которые ссылается данная страница. Множество входящих ссылок содержит страницы из множества V, которые ссылаются на данную страницу.

Мера схожести двух веб-страниц $i, j \in V$ на основе исходящих ссылок может быть определена в виде

$$Sout(i, j) = \frac{|out(i) \cap out(j)|}{|out(i) \cup out(j)|} \qquad (1)$$

Мера $Sout(i, j)$ показывает, как много общих ссылок содержат страницы i и j, в отношении к совокупному количеству ссылок на данных страницах. Значения $Sout(i, j)$ принадлежат целочисленному интервалу $[0,1]$. Мера принимает значение равное 0, если нет ни одной общей исходящей ссылки для данных двух страниц. Если же все ссылки на данных двух странницах идентичны, то мера принимает значение равное 1. Если оба множества $out(i)$ и $out(j)$ пустые, то мера определяется равной 0.

Подобным же образом может быть определена мера схожести двух страниц $i, j \in V$ на основе входящих ссылок

$$Sin(i, j) = \frac{|in(i) \cap in(j)|}{|in(i) \cup in(j)|}. \quad (2)$$

Данная мера обладает теми же свойствами, что и предыдущая. Она показывает насколько много страниц ссылается как на i, так и на j, в отношении к совокупному количеству страниц, которые ссылаются на эти две страницы. Если оба множества $in(i)$ и $in(j)$ пустые, то мера определяется равной 0.

Теперь мы можем ввести понятие меры схожести веб-страниц как линейной комбинации мер схожести на основе входящих и исходящих ссылок

$$S(i, j) = \alpha Sout(i, j) + (1-\alpha) Sin(i, j), \quad (3)$$

где $i, j \in V$, а α является варьируемым параметром и принадлежит интервалу $[0,1]$. Параметр α определяет степень вклада входящих и исходящих ссылок в меру ссылочной схожести двух страниц. Значения меры принадлежат целочисленному интервалу $[0,1]$. Мера принимает значение 0, если страницы абсолютно не похожи в смысле ссылочной схожести и 1, если они идентичны. Стоит заметить, что интерпретация значений данной меры зависит от выбранного значения параметра α.

После определения схожести двух страниц, на основе матрицы A_{ij} может быть построена матрица схожести страниц

$$S_{ij} = S(i, j), \forall i, j \in V. \quad (4)$$

Данная матрица является симметричной относительно главной диагонали и определяет схожесть всех веб-страниц принадлежащих графу G. Данная матрица может рассматриваться как матрица смежности полного взвешенного неориентированного графа $GS = < VS, ES >$ – графа схожести

страниц, где VS множество веб-страниц, а ES множество взвешенных ребер. Весом ребра $< i, j >$ является мера схожести двух страниц $S(i, j), \forall i, j \in VS$.

Стоит заметить, что существует возможность динамического варьирование значения параметра α в зависимости от значений мер $Sout(i, j)$ и $Sin(i, j)$. Это может быть использовано для обнаружения разных типов ссылочного спама.

Кластеризация на основе графа схожести страниц

Для достижения поставленной цели, нахождения кластеров схожих страниц, можно воспользоваться алгоритмом выделения связных компонент графа.

Задав пороговой параметр схожести $R \in [0,1]$, мы можем удалить из множества ES все ребра $< i, j >$, для которых вес $S(i, j) < R$. После удаления этих ребер исходный граф разделится на несколько связных компонент, которые и будут искомыми кластерами. Связные компоненты в модифицированном графе можно найти одним из классических методов теории графов.

Достоинством метода служит то, что могут быть найдены кластеры произвольно конфигурации. Недостатком данного метода является зависимость от параметра пороговой схожести, который необходимо задавать вручную.

Значение параметра R может быть подобрано эмпирически после анализа результатов разбиения на тестовой выборке.

Заключение

В данной статье введено понятие ссылочной схожести веб-страниц и предложен простой алгоритм выделения кластеров страниц, которые могут быть подозрительны с точки зрения использования ссылочного спама. Данный метод планируется проверить на основе данных собранных с главных страниц веб-сайтов, входящих в украинский сегмент Интернет (около 100000 единиц). Также необходимо изучить влияние параметра пороговой схожести R на результаты кластеризации и возможность динамической вариации параметра α. Возможность использования предложенного алгоритма для динамической фильтрации результатов поиска также требует проверки.

Данная работа была проделана при поддержке ЗАО «МЕТА» (http://www.meta.ua). Дальнейшие эксперименты планируется проводить как с использованием данных, предоставленных этой компанией, так и данных, которые возможно открыто получить от других компаний, разрабатывающих поисковые сервисы.

Список литературы: ***1.*** *Baoning Wu and Brian D. Davison.* Identifying link farm spam pages. – In Proc. of the 14th International WWW Conference, 2005. ***2.*** *Z. Gyöngyi and H. Garcia-Molina.* Web spam taxonomy. – In First International Workshop on Adversarial Information Retrieval on the Web, 2005. ***3.*** *Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd*. The PageRank citation ranking: Bringing order to the web. – Technical report, Stanford University, 1998. ***4.*** *J. M. Kleinberg.*

Authoritative sources in a hyperlinked environment. – In proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms (SODA), 1998. **5.** *Z. Gyöngyi and H. Garcia-Molina.* Link spam alliances. – In Proceedings of the 31st International Conference on Very Large Data Bases (VLDB), 2005. **6.** *da Costa-Carvalho, A. L., Chirita, P.-A., de Moura, E. S., Calado, P., and Nejdl, W.* Site level noise removal for search engines. – In Proceedings of the 15th international conference on World Wide Web, 2006. **7.** *G. Roberts and J. Rosenthal*. Downweighting tightly knit communities in world wide web rankings. – Advances and Applications in Statistics (ADAS), 2005. – №. 3 – С. 199-216

Поступила в редколлегию 16.05.07

УДК 519.854.2

О.А.ПАВЛОВ, д-р. техн. наук, НТУУ «КПІ»,
О.Б.МІСЮРА, канд. техн. наук, НТУУ «КПІ»,
О.А.ХАЛУС, асп. НТУУ «КПІ»

ЗАДАЧА МІНІМІЗАЦІЇ СУМАРНОГО ЗАПІЗНЕННЯ ВИКОНАННЯ НЕЗАЛЕЖНИХ ЗАВДАНЬ З ДИРЕКТИВНИМИ СТРОКАМИ ОДНИМ ПРИЛАДОМ В СИСТЕМІ ПЛАНУВАННЯ ТА УПРАВЛІННЯ ДРІБНОСЕРІЙНИМ ВИРОБНИЦТВОМ (СПУДВ)

В статті розглянута задача мінімізації сумарного запізнення виконання незалежних завдань з директивними строками одним приладом, яка входить до складу математичного забезпечення системи СПУДВ. Ця задача відноситься до NP- складних, що обумовлює складність пошуку не тільки точних методів розв'язання задачі, але і наближених. Запропоновано ефективний точний ПДС-алгоритм (алгоритм із поліноміальною й експоненційною складовими) розв'язання задачі, заснований на новому підході до розв'язання задач з директивними строками, що полягає в оптимальному використанні резервів часу незапізнених завдань.

Вступ

Задача сумарного зваженого запізнення привертає увагу вчених протягом багатьох років, але через надмірні обчислювальні вимоги точне розв'язання для задачі з 50 завданнями – це бар'єр, що практично неможливо перебороти.

Запропоновано новий підхід до розв'язання задачі і розроблений на його основі алгоритм, що дозволяє одержувати точні розв'язки для задач з числом змінних $n > 500$ [1, 2]. Цей підхід заснований на конструктивної теорії розв'язання важкорозв'язуємих задач комбінаторної оптимізації, розробленої під керівництвом професора О.А. Павлова [3]. Основна ідея теорії полягає в дослідженні властивостей розглядуваних класів важкорозв'язуємих задач, доказі положень, правил, що дозволяють розробити єдиний принцип обчислень, і на їх основі побудові ПДС-алгоритмів.