РАСШИРЕНИЕ ПРЕДСТАВЛЕНИЯ ДОКУМЕНТОВ ПРИ ПОИСКЕ В ВЕБ[1]

Владимир Иванов, Игорь Некрестьянов, Надежда Пантелеева
Санкт-Петербургский Государственный Универсиетет, 198504, Россия,
Санкт-Петербург, Старый Петергоф, Университетский пр. 28
e-mails: 1085@mail.ru, igor@meta.math.spbu.ru, nadejda@meta.math.spbu.ru

В работе исследуется возможность использования информации о содержимом документов в окрестности рассматриваемых страниц Веб для повышения качества результатов поиска на примере задачи классификации.

EXPANDING A WEB DOCUMENT REPRESANTATION

Vladimir Ivanov, Igor Nekrestyanov, Nadejda Panteleeva
Saint-Petersburg State University, Universitetsky pr, 28, St.Petergoff,
St.Petersburg, 198504, Russia

The idea of changing the granularity level of Web search from one Web page to Web page and it's nearest neighborhood is not new and seems to be quite useful in many of the Web Information Retrieval tasks. This paper studies the influence of the implementation of such idea on the quality of Web page classification.

1 Введение

Многие современные поисковые системы в Веб основаны на хорошо известных методах и алгоритмах, разработанных еще до появления Интернет [7, 23].

Однако, хотя в общем виде постановки многих задач поиска в Веб не сильно отличаются от традиционных, Веб имеет ряд специфических особенностей, которые зачастую влекут некоторое смещение постановки поисковой задачи, и которые необходимо учитывать при ее решении [3].

Так в наиболее традиционной постановке задачи информационного поиска идет речь о поиске документов, которые содержат необходимую пользователю информацию. В классической интерпретации поиск производится по плоским текстовым файлам, каждый из которых содержит по одному документу. Таким образом, в этом случае речь идет о гранулярности поиска на уровне файлов.

[1] Эта работа частично поддержана грантом РФФИ 01-01-00935.

Однако в некоторых ситуациях имеет смысл изменение гранулярности. Типичным примером уменьшения гранулярности поиска является рассмотрение отдельных параграфов в качестве единиц информации [7]. В контексте Веб разбиение документа на части используется для достижения разных целей - например, как для построения более точных профилей документа [4], так и для уточнения графа при его анализе в алгоритме ранжирования HITS [8].

Другой интересный пример уменьшения гранулярности - задача фактографического поиска, где необходимо обнаружить конкретный ответ на вопрос пользователя, например, «В каком году состоялась Куликовская битва?» [3, 5, 19].

Однако в рамках этой работы нас больше интересует увеличение гранулярности. Действительно, зачастую логически единые документы в гипертекстовых системах представляются набором связанных гиперссылками страниц. Более того, известно, что Веб обладает свойством тематической локальности [9].

Основной целью этой работы является проверка гипотезы о том, что использование информации о других страницах Веб из окрестности рассматриваемой страницы может быть использовано для повышения качества результата поиска. Естественно, что дополнительная информация может привносить и дополнительный шум, поэтому мы также хотим установить, какие страницы стоит учитывать, а какие - нет.

Для проведения количественных оценок мы рассматривали влияние этой дополнительной информации на задачу тематической классификации и использовали стандартные критерии оценки качества классификации [2]. Кроме этого мы также рассмотрели множество других дополнительных объективных характеристик в надежде определить зависимость между этими характеристиками и вкладом дополнительной информации в качество классификации.

Статья организована следующим образом: в следующем разделе мы охарактеризуем известные нам исследования на близкие темы; в разделе 3 мы опишем рассматриваемые нами методы уточнения представления документов, а в разделе 4 представим результаты их экспериментальной оценки.

2 Близкие работы

Особенности Веб и пользователей информационно-поисковых систем в Веб обуславливают проведение новых исследований, направленных на повышение качества работы различных методов информационного поиска в приложении к Веб [3].

2.1 Использование информации о контексте

Одним из активно исследуемых источников дополнительной информации является контекст, в котором выполняется поиск.

Говоря о контексте поиска, чаще всего имеют в виду нечто, что позволяет дополнительно характеризовать решаемую задачу поиска, например, информацию о текущих интересах автора поискового запроса [20].

В этом случае обычно накапливают информацию о поведении пользователя - какие страницы из найденных он посещал, сколько времени он проводил, просматривая отдельные страницы, и т.п. Целью подобных исследований является разработка алгоритмов персонализации, которые автоматически уточняют результаты поиска под потребности конкретного пользователя [3].

Другой подход к преодолению сложностей, вызванных «бедностью» типичных запросов пользователей - использование методов автоматического расширения запросов. Например, для расширения запросов методом *анализа локального контекста (LCA)* используется информация о контексте, в котором встречаются термы запроса в реальных документах [28].

Еще одна идея - использовать информацию о контексте, в котором пользователь нашел тот набор ключевых слов, который он использовал для задания запроса, - реализована в системе InteliZap [10]. Поскольку зачастую потребность в поиске возникает в процессе чтения электронных источников для получения дополнительных подробностей, то такой подход позволяет четче понять, что же реально интересует пользователя.

Однако возможно и использование информации о контексте информационных ресурсов. Например, в работе [6] текст в ссылках на страницу использовался для уточнения информации об этой странице в задаче кластеризации. Подобный подход использовался и для определения «репутации» ресурсов Веб в работе [22]. Дополнительную информацию можно получить и рассматривая ссылку на ресурс как своеобразную текстовую строку [24].

В контексте Веб можно использовать информацию о структуре Веб (или регулярностях этой структуры) для повышения качества классификации [12, 13].

2.2 Тематическая локальность

Интересной особенностью Веб является *тематическая локальность*, т.е. корреляция между отличием в тематике содержания страниц и расстоянием в графе Веб. Эмпирическое подтверждение этой интуитивно кажущейся естественной закономерности было получено сравнительно недавно [9].

Более формально эмпирически показана справедливость следующих предположений:

- Ссылки в большинстве случаев соединяют страницы с близким содержанием
- Текст ссылки и, возможно, вокруг нее описывает содержание страницы, на которую ведет ссылка.

Эта особенность Веб активно используется при разработке эффективных стратегий сканирования для сетевых роботов, а также при решении некоторых других задач - выделении тематических «сообществ» [11], алгоритмах ранжирования [18] и т.д.

2.3 Логические документы

Идея повысить уровень гранулярности единицы информации, т.е. перейти от взаимосвязи «*один файл - один документ»* к «*много связанных гиперссылками файлов - один документ»* и, тем самым рассматривать «логические» документы, не нова [25].

За последние несколько лет был предложен ряд подходов к выделению логических документов в Веб, которые можно разделить на статические и динамические [27,14].

В рамках первого подхода множество документов поиска заранее делится на некоторые логические документы, которые и используются при обработке запросов [27]. Достоинством этого подхода является возможность использования более трудоемких методов определения логических документов, поскольку этот процесс не сказывается на времени обслуживания запросов. Однако гарантировать правильное разбиение невозможно (на практике оно очень часто вовсе не очевидно даже автору ресурса), и ошибки в разбиении понижают эффективность системы.

Альтернатива - выделять логические документы динамически. Поскольку нам необходимы только те логические документы, которые удовлетворяют условию поиска, то и определять достаточно только их границы, причем не обязательно абсолютно точно - достаточно обнаружить их часть, удовлетворяющую запросу [26].

Для выявления логических единиц могут быть использованы различные источники информации, например, содержание страниц (схожесть лексики, стиля), структура связывающих их ссылок. Так структура каталогов файловой системы обычно отражает представления автора о логической структуре, и поэтому анализ ссылок на документы полезен для определения этой структуры.

В работе [14] показано, что использование логических документов вместо отдельных Веб страниц повышает точность кластеризации результатов поиска. А в работе [26] логические документы применялись как единицы информации в процессе поиска, т.е. ответом на запрос является ссылка на логический документ, который содержит все термы запроса, но они не обязательно встречаются в какой-либо его одной странице.

Несмотря на обнадеживающие первые результаты, исследование применимости концепции логических документов к задачам поиска в Веб на настоящий момент носит фрагментарный характер и требует проведения дополнительной работы в этом направлении.

3 Методы представления документа Веб

Традиционно представления Веб страниц строятся по той информации, которая содержится в самих этих страницах, что следует подходам, принятым в классических текстовых ИПС [7].

Однако гиперссылки, которые связывают страницы в Веб, зачастую не только предоставляют механизм навигации, но также зачастую характеризуют семантические взаимосвязи между страницами.

Поэтому вполне логичной кажется идея использовать информацию об окружении конкретной страницы Веб *p* для более точного ее представления при решении задач информационного поиска.

В этой работе мы рассматриваем несколько возможных подходов к расширению представления документа *p*:

«Базовое» (*Base*).

В этом случае представление документа строится по единственной Веб странице (собственно странице *p*).

«Наивное расширение» (*Greedy*).

Представление строится на основе объединения содержимого всех документов, которые находятся на расстоянии в не более, чем *N* ссылок от *p*.

Поскольку используемые в Веб ссылки - однонаправленные, то реальная материализация такой окрестности для произвольной Веб страницы в большинстве случаев исключительно трудоемка, и поэтому мы ограничиваемся рассмотрением только исходящих путей.

«Общий сервер» (*SameServer*).

Сужение предыдущего представления за счет исключения всех документов, которые расположены на другом Веб сервере, чем страница *p*.

«Общая директория» (*SameDir*).

Дальнейшее сужение, при котором рассматриваются только те страницы, которые расположены в той же директории Веб сервера, что и *p*.

«Модифицированное» (*Modified*)

Представление, при построении которого частично использовался ручной анализ для принятия решения об исключении страниц из рассмотрения.

Отметим, что такое представление является подмножеством представления *Greedy*.

При создании представлений производилась предварительная обработка страниц, заключавшаяся в удалении разметки HTML и приведении используемой кодировки к общему виду.

3.1 Рабочие гипотезы

Интуитивно кажется, что тематическая локальность Веб должна повлечь увеличение объема полезной информации, которая характеризует страницу при расширении ее представления. Эта идея и лежит в основе первой из проверявшихся гипотез.

Гипотеза 1.

Расширение представления позволяет повысить качество решения задач поиска.

С другой стороны, тематическая локальность Веб не означает, что любые две связанные ссылкой страницы относятся к одной и той же тематике, поэтому весьма вероятно, что при расширении в построенное представление попадет мусор. Это соображение обуславливает вторую из рассматриваемых гипотез.

Гипотеза 2.

Прямолинейные методы расширения зачастую излишне повышают уровень шумов в представлении и поэтому дают не идеальный результат.

Отметим, что эти гипотезы не являются противоположными, а дополняют друг друга.

4 Экспериментальная оценка

Целью проводимых экспериментов является подтверждение или опровержение рабочих гипотез, которые были сформулированы в предыдущем разделе.

4.1 Постановка эксперимента

Нам не известен ни один стандартный тестовый набор данных для классификации, который бы состоял из русскоязычных страниц Веб. Поэтому мы использовали набор данных, построенный на базе каталога ресурсов Веб - *List.Ru* (http://list.ru).

Поскольку страницы для тематических категорий в этом каталоге отбираются экспертами, то эти категории можно считать эталонными результатами при оценке качества автоматической классификации тех же страниц Веб.

Используемый нами набор данных был построен на основе 50 выбранных вручную тематических категорий второго и третьего уровня. Из каждой категории мы взяли первые 100 ссылок. После исключения некорректных[2] ссылок итоговый набор $Dataset_{50}$ содержал 4734 ссылки.

[2] К некорректным относились ссылки, которые нам по каким-либо причинам не удалось скачать.

	размер представления		
	слов	**Kb**	**физических страниц**
BASE	1661032	29620	4734
GREEDY	67093098	718920	135274
SAME-SERVER	52351561	544724	98589
SAMEDIR	17413087	186400	32753

Таблица 1. Характеристики представлений для набора $Dataset_{50}$

При проведении экспериментов мы также использовали сужения $Dataset_k$ полного набора по количеству категорий (т.е. суженный набор содержал только ссылки, которые относились к одной из k выбранных категорий).

При проведении эксперимента набор данных разбивался на 2 части: тренировочное множество, которое использовалось для обучения классификатора, и множество для тестирования - в соотношении 0.6 к 0.4 соответственно.

Для того чтобы нивелировать погрешность, вызванную спецификой конкретного разбиения или построения суженного набора данных, все эксперименты повторялись по несколько раз, а их результаты усреднялись. При этом каждый раз документам для обучения/тестирования в *Base* соответствовали их расширенные версии для обучения/тестирования в *SameDir*, *SameServer*, и *Greedy*.

В таблице 1 представлена сводная статистическая информация о построенных представлениях для набора $Dataset_{50}$ (см. раздел 3). Информация для суженных представлений может быть легко получена путем соответствующего линейного масштабирования значений в таблице.

4.2 Влияние на классификацию

Целью этой серии экспериментов является проверка гипотезы о потенциальной полезности использования расширенных форм представления информации о Веб странице при решении задачи тематической классификации.

4.2.1 Методы классификации

За годы исследований в области автоматической классификации текстовой информации было предложено множество различных методов классификации, а также разнообразных модификаций принципов выбора и взвешивания признаков, которые используются для представления текстовой информации [1, 21, 16, 15, 17]. Тем не менее невозможно выделить единый оптимальный способ решения этой задачи.

В рамках этой работы мы не претендуем на попытку улучшить алгоритмы классификации и, поэтому используем известные популярные подходы. В частности, мы рассматриваем следующие методы классификации[3]:

- **метод Байеса** (NB) [16]

Согласно этому методу документ *d* считается принадлежащим наиболее вероятной категории C_k:

$$k = argmax_j P(C_j|d)$$

Условные вероятности $P(C_j|d)$ вычисляются на основе тренировочного набора данных.

- **метод опорных векторов** (SVM) [21], [17]

В этом методе вероятность ошибки при классификации выбранного произвольно и ранее не встречавшегося документа минимизируется посредством разделения пространства классифицируемых документов на области, соответствующие различным классам документов. Разделители описываются функцией-ядром. В наших экспериментах использовались функции-ядра 1-го порядка, т.е. классы определялись набором гиперплоскостей.

- **TFIDF** [16]

Здесь определение класса документа $D(C_k)$ происходит по следующей формуле:

$$C_k = argmax_{Cj \in \mathbf{C}} \cos(\mathbf{c}(C_j), \mathbf{d}`),$$

где **d`** - вектор, каждая компонента которого отвечает за вес (TF*IDF) соответствующего по номеру слова в документе **d**. Зависимость компонент вектора $\mathbf{c}(C_j)$ от класса C_j определяется в процессе обучения.

- **метод вероятностного индексирования** (prind) [29]

Этот метод основывается на предположении того, что характерные термы различных категорий (по результатам обучения) с разной вероятностью распределяются в тестовых документах.

4.2.2 Критерий оценки качества

Для того чтобы сравнить качество классификации мы использовали критерий *аккуратности* (*Accuracy*) классификации, который вычисляется следующим образом:

---

[3] При проведении экспериментов мы использовали свободно распространяемую реализацию этих методов в пакете *Rainbow*

(http://www-2.cs.cmu.edu/~mccallum/bow/rainbow/).

$$Accuracy = N_{correct} / N_{incorrect} * 100\%$$

где $N_{correct}$ - количество правильно классифицированных документов,

$N_{incorrect}$ - количество неправильно классифицированных документов.

4.2.3 Результаты

Результаты экспериментов представлены на рис. 1.

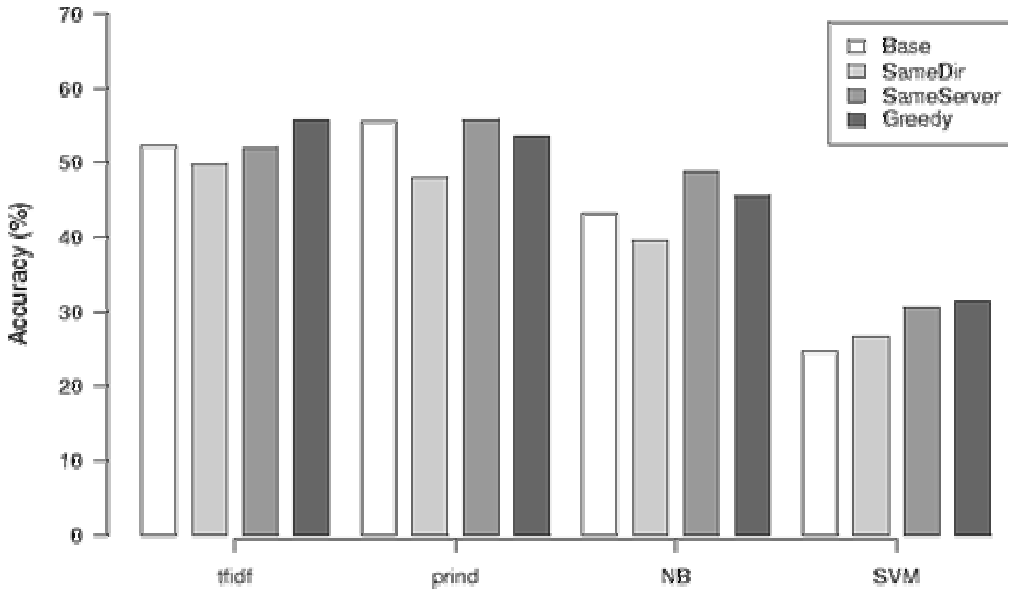

Рисунок 1. Аккуратность классификации для разных представлений документа набора *Dataset*$_{50}$

Использование какого-либо одного из трех расширенных представлений документа дало заметный прирост в аккуратности классификации всеми методами, что полностью согласуется с гипотезой 1. Использование представления *SameDir* повлекло падение качества классификации почти всеми методами по сравнению с базовым представлением, также использование *Greedy* понизило аккуратность классификации по сравнению с *SameServer* методами Байеса и *prind*. Этот факт, по нашему мнению, хорошо согласуется с гипотезой 2 - прямолинейные подходы к расширению представления влекут наряду с увеличением полезной информации также увеличение и бесполезной (шумовой) информации в различном для кажлого подхода соотношении. Более того, тот факт, что для разных методов классификации наблюдается различный порядок изменения качества классификации при расширении представления документа, говорит о том, что разные методы классификации по-разному учитывают соотношение полезной и шумовой информации в документе. Поэтому одно расширенное представление эффективно для одного метода классификации, другое - для другого. В экспериментах на множестве $Dataset_2$ мы использовали представление *Modified*, которое, несмотря на незначительный объем ручной обработки, всегда давало лучший результат по сравнению с другими расширенными представлениями. Отметим, что

ширенными представлениями. Отметим, что обработка *Modified* заключалась в частичном удалении Веб страниц из расширенного представления, которые, очевидным образом, не принадлежат логическому документу базовой страницы и являются шумовой информацией.

Этот результат показывает перспективность исследования автоматических подходов к расширению представления документа, которые бы могли контролировать вклад шумовой информации в отличие от используемых нами прямолинейных подходов.

4.2.4 Оценка стабильности результатов

Нам было интересно узнать, подтвердится ли гипотеза 1 и соответствующий результат наших экспериментов при изменении соотношения документов для обучения и для тестирования в наборе $Dataset_{50}$ (см. таблицу 2).

Мерой стабильности/нестабильности результата мы выбрали количество классификаций, когда результат подтверждался, опровергался или был не сравним. При этом мы считали, что результат подтверждается/опровергается, если аккуратность классификации при использовании расширенного представления выше/ниже более, чем на 5%, по сравнению с базовым представлением (значения 1/-1 в таблице 2). В противном случае мы считали, что результат не сравним (значение 0 в таблице 2). В качестве расширенного представления мы взяли *SameServer*, а в качестве метода классификации - метод Байеса. Для каждого заданного соотношения разбиения (строки таблицы 2) мы сделали 5 случайных разбиений (столбцы таблицы 2) и на этих разбиениях сравнили аккуратности классификации. При этом каждый раз документам для обучения/тестирования в *Base* соответствовали их расширенные версии для обучения/тестирования в *SameServer*. По данным, представленным в таблице 2, видно, что, если доля документов для обучения больше половины, то гипотеза 1 подтверждается абсолютно стабильно.

Случаев, когда гипотеза 1 опровергалась, мы не зафиксировали.

доля документов	**номер случайной выборки**				
для обучения	**1**	**2**	**3**	**4**	**5**
0.2	0	0	0	0	0
0.4	0	0	0	0	0
0.6	1	1	1	1	1
0.8	1	1	1	1	1

Таблица 2. Стабильность результата относительно соотношения документов для обучения и для тестирования

кол-во	номер случайной выборки				
категорий	1	2	3	4	5
2	0	-1	1	1	1
5	1	1	1	1	0
10	0	1	1	1	1
25	1	1	1	1	1

Таблица 3. Стабильность результата относительно количества участвующих категорий

	размер	
	словаря (слов)	**пересечения test/train (%)**
Base	229073	25
SameDir	789573	26,5
SameServer	1344656	28
Greedy	1494319	28,31

Таблица 4. Характеристики словарей представлений для набора $Dataset_{50}$

Аналогично, мы оценили стабильность результата относительно изменения количества категорий в наборе данных. В таблице 3 строки соответствуют набору данных (количеству участвующих категорий), а столбцы - номеру случайной выборки данного количества категорий из $Dataset_{50}$. Как видно по данным таблицы 3 стабильность результата снижается с уменьшением количества категорий.

4.4.4 Изменение объективных характеристик

Для того, чтобы лучше обосновать наблюдаемое поведение аккуратности классификации, мы попытались собрать и проанализировать статистику о некоторых объективных характеристиках рассматриваемых представлений (частично эта информация собрана в таблице 4).

Целью этого анализа является попытка разобраться в том, изменение каких объективных характеристик сопутствует улучшению представления Веб страницы при ее расширении.

Для этого, мы рассмотрели пересечение словарей представлений документов для тестирования и словарей, построенных по представлениям документов той же категории, но использовавшихся для обучения, в соотношении 0.4 к 0.6 соответственно. При переходе от базового к расширенным представлениям как абсолютный, так и относительный размеры пере-

сечения возрастали. Такое наблюдение хорошо коррелирует с рабочей гипотезой о повышении объема доступной информации.

5 Заключение

В рамках этой работы мы исследуем возможность использования информации о содержимом документов в окрестности рассматриваемых страниц Веб для повышения качества результатов поиска на примере задачи классификации.

Результаты проведенных экспериментов подтверждают обе сформулированные базовые гипотезы - расширение информации о документе позволяет повысить качество классификации, но применение прямолинейных подходов к расширению не позволяет достичь идеального результата. Тем самым мы показываем перспективность исследования более сложных методов расширения.

Оценка стабильности полученных результатов относительно соотношения количества документов для обучения и тестирования классификатора показывает, что всегда имеет место улучшение качества классификации расширенного представления документа, и оно более заметно с увеличением доли документов для обучения в коллекции. Однако с уменьшением количества категорий для классификации результаты оказываются менее стабильными.

В дальнейшем мы собираемся рассмотреть эффективность применения обучаемых подходов к автоматическому расширению. В частности, перспективной кажется идея расширения только за счет страниц из того же логического документа.

Литература

1. И.Е. Кураленок, И.С. Некрестьянов. Автоматическая классификация документов с использованием семантического анализа // Программирование, 4:31-41, 2000.
2. И. Кураленок, И. Некрестьянов. Оценка систем текстового поиска // Программирование (в печати), 2002.
3. И. Некрестьянов, Н. Пантелеева. Системы текстового поиска для Веб // Программирование (в печати), 2002.
4. А. Г. Дубинский. Разработка моделей и совершенствование структуры систем информационного поиска в глобальной компьютерной сети // PhD thesis, ДНУ, 2002.
5. Eugene Agichtein, Steve Lawrence, and Luis Gravano. Learning search engine specific query transformations for question answering. In *Proc. of the WWW10*, pages 169-178, 2001.

6. Giuseppe Attardi, Antonio GullM, and Fabrizio Sebastiani. Automatic Web page categorization by link and context analysis. In Chris Hutchison and Gaetano Lanzarone, editors, *Proc. of the THAI-99*, pages 105-119, Varese, IT, 1999.
7. Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, 1999.
8. Soumen Chakrabarti. Integrating the document object model with hyperlinks for enhanced topic distillation and information extraction. In *Proc. of the WWW10*, pages 211-220, 2001.
9. Brian D. Davison. Topical locality in the web. In *Proc. of the SIGIR'00*, pages 272-279, 2000.
10. Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. Placing search in context: the concept revisited. In *Proc. of the WWW10*, pages 406-414, 2001.
11. Gary Flake, Steve Lawrence, and C. Lee Giles. Efficient identification of web communities. In *Proc. of the SIGKDD'00*, pages 150-160, August 2000.
12. Rayid Ghani, SeAn Slattery, and Yiming Yang. Hypertext categorization using hyperlink patterns and meta data. In *Proc. of ICML-01*, pages 178-185, 2001.
13. Eric Glover, Kosas Tsioutsiouliklis, Steve Lawrence, David Pennock,and Gary Flake. Using web structure for classifying and describing web pages. In *Proc. of the WWW'2002*, May 2002.
14. Kenji Hatano, Ryouichi Sano, Yiwei Duan, and Katsumi Tanaka. An interactive classification of web documents by self-organizing maps and search engines. In *Proc. of the DASFAA'99*, pages 35-42, 1999.
15. Ufuk Ilhan. Application of K-NN and FPTC based text categorization algorithms to Turkish news reports. Master's thesis, Computer Engineering Department, Bilkent University, 2001.
16. Thorsten Joachims. A probabilistic analysis of the rocchio algorithm with TFIDF for text categorization. In *Proc. of the ICML'97*, pages 143-151, 1997.
17. Thorsten Joachims. Text categorization with support vector machines: learning with many relevant features. In *Proc. of ECML-98*, number 1398, pages 137-142, 1998.
18. Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604-632, 1999.
19. Cody T. Kwok, Oren Etzioni, and Daniel S. Weld. Scaling question answering to the web. In *Proc. of the WWW10*, pages 150-161, May 2001.
20. Steve Lawrence. Context in web search. *IEEE Data Engineering Bulletin*, 23(3):25-32, 2000.
21. Edgar Osuna, Robert Freund, and Federico Girosi. Support vector machines: Training and applications. Technical Report AIM-1602, 1997.

22. Davood Rafiei and Alberto Mendelzon. What is this page known for? computing web page reputations. In *Proc. of the WWW9*, pages 823-835, May 2000.
23. G. Salton and M. J. McGill. *Introduction to modern Information Retrieval*. McGraw-Hill Computer Science Series. McGraw-Hill, New York, 1983.
24. Vojtech Svatek and Petr Berka. URL as starting point for WWW document categorisation. In *Proc. of the RIAO'2000.*, pages 1693-1702, 2000.
25. Keishi Tajima, Kenji Hatano, Takeshi Matsukura, Ryouichi Sano,and Katsumi Tanaka. Discovery and retrieval of logical information units in web. In *Proc.of the WOWS'99*, August 1999.
26. Keishi Tajima, Kenji Hatano, Takeshi Matsukura, Ryouichi Sano,and Katsumi Tanaka. Discovery and retrieval of logical information units in web. In *Proc. of ACM DL'99*, pages 13-23, August 1999.
27. Keishi Tajima, Yoshiaki Mizuuchi, Masatsugu Kitagawa, and Katsumi Tanaka. Cut as a querying unit for WWW, Netnews, and E-mail. In *Proc. of Hypertext'98*, pages 235-244, June 1998.
28. Jinxi Xu and W. Bruce Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems*, 18(1):79-112, 2000.
29. Yiming Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. In *International Conference on Machine Learning*, pages 412-420, 1997.