Российская академия наук
Сибирское отделение
Институт систем информатики
им. А. П. Ершова

А.В. Будний, А.Д. Русанов

МЕТОД ПРОФИЛИРОВАНИЯ ПОЛЬЗОВАТЕЛЕЙ ИНТЕРНЕТ-ПОРТАЛОВ СРЕДСТВАМИ SEMANTIC WEB

Препринт
151

Новосибирск 2009

Профилирование – это разумное ограничение предъявляемой посетителю информации с целью выделения более важного для него содержания. Задача, стоящая перед авторами, заключается в разработке нового метода профилирования пользователей Интернет-порталов и реализации этого метода в виде веб-сервера документов, «предугадывающего» предпочтения посетителя.

На основе ряда признаков из всех понятий онтологии выделяются наиболее общие, называемые тематиками. Каждому понятию онтологии сопоставляем вектор, каждая координата которого характеризует близость понятия к одной из тематик. На основе совокупности векторов понятий, упомянутых в документе, строится вектор, характеризующий близость всего документа к каждой из тематик, называемый профилем документа. Каждый посетитель обладает набором интересов, и его тоже можно описать подобным вектором, называемым профилем посетителя. Однако если профиль документа является статичным, вектор интересов посетителя постоянно корректируется с просмотром новых документов и изменением интересов пользователя со временем. Наконец, чтобы понять, какие документы более интересны пользователю на текущий момент, необходимо соотнести профиль пользователя с профилями документов. Степень соответствия документа интересам пользователя определяется углом между этими двумя векторами и их длинами. Неплохой оценкой близости является скалярное произведение векторов.

Ключевые слова: профилирование пользователей, Semantic Web, семантические сети, векторная модель информационного поиска.

Siberian Division of the Russian Academy of Sciences
A. P. Ershov Institute of Informatics Systems

A.V. Budniy, A.D. Rusanov

METHOD OF INTERNET PORTAL VISITORS PROFILING USING SEMANTIC WEB

Preprint
151

Novosibirsk 2009

Visitor profiling is a reasonable server content filtering method, which allows to provide content according to the visitor's preferences. The authors' goal is developing a new approach to Internet portal visitors profiling and implementing this method in form of a document web-server capable of forecasting the visitor's interests.

The main notions in an ontology are called subjects, they are extracted from the ontology based on certain criteria. Each ontology notion has a corresponding vector, whose coordinates characterize the closeness of the notion to a subject. It is possible to form a vector characterizing the closeness of a document to each subject based on the vectors of notions mentioned in the document. This vector is called a document profile. Each visitor has interests and therefore can be described with a similar vector, called a visitor profile. However, in contrast to the document vector, which is static, the visitor vector is adjusted each time the visitor reads a new document. Therefore, the visitor profile changes as the visitor's interests change. The visitor profile is compared with document profiles to find out which documents are more interesting to the user at the moment. The goodness is determined by the angle between the two vectors and their length. We can get quite accurate estimation using simple scalar vector multiplication.

Keywords: Visitors profiling, Semantic Web, semantic networks, vector model of information retrieval, ontology

ВВЕДЕНИЕ

Ежедневно пользователи всемирной паутины получают массу информации. Иногда она бывает полезной, однако большая ее часть не представляет интереса для человека, который ее просматривает. Это связано с тем, что каждый человек – это личность, он уникален. То, что представляется занимательным одному, другому может показаться бессмысленным. Поэтому у каждой статьи найдется благодарный читатель, однако вероятность того, что им будет именно тот, кому она в данный момент предъявлена, очень невелика.

Для того чтобы изменить такое положение дел, необходимо обдуманно подойти к тому, какую информацию показывать конкретному пользователю, вместо того, чтобы действовать наугад или полагаться исключительно на предпочтения редактора, как это часто происходит. Чтобы принять взвешенное решение о выборе данных для отображения, необходима информация о потребителе, которая позволила бы сделать вывод о его предпочтениях и интересах, либо помогла по другим признакам отсечь информацию, которая заведомо не представляет ценности для данного Интернет-пользователя.

Профилирование пользователя – достаточно известная и распространенная задача, решаемая в настоящее время различными способами [7, 18]. Суть профилирования заключается в том, что пользователю произвольного информационного Интернет-ресурса предоставляют не весь контент, а в первую очередь то, в чем, предположительно, он может быть заинтересован. Предположение обычно строится на основе многих факторов: документов, которые пользователь смотрел в прошлом, его географического положения, приватной информации из личного профиля пользователя и т.д. В данной работе для решения задачи профилирования определяются тематики, интересные пользователю, и оценивается близость того или иного документа к необходимым тематикам.

В работе описан оригинальный способ решения задачи профилирования пользователя веб-сервера документов, основанный на использовании семантических сетей [1, 2, 10].

Постановка задачи

Целью работы является создание и проверка предложенных методов профилирования Интернет-пользователей. Испытание новых методик пла-

нируется проводить на базе веб-сервера документов, который каждому пользователю предоставляет индивидуальный набор документов в соответствии с его интересами.

В большинстве случаев документы на сервере упорядочены по темам, по алфавиту или иным образом, однако всем пользователям данные представляются одинаково, без учета их персональных особенностей. Это приводит к тому, что посетитель тратит больше времени на поиск нужной ему информации.

Более перспективной идеей видится организация контента в соответствии с интересами пользователя, на основе ранее просмотренных документов. Тогда посетитель, входя на сайт, видит не весь список документов, а в первую очередь те документы, которые именно ему могут быть более интересны.

Если человек в прошлом интересовался какой-то темой и читал соответствующие документы на сервере, то есть смысл предложить ему в первую очередь похожие документы.

Опишем в общих чертах реализацию сервера документов. Более детально все будет рассмотрено в последующих разделах статьи.

Чтобы иметь возможность сказать, какие документы похожи, перед сохранением на сервер каждый документ проходит специальный анализ, в ходе которого выделяются основные темы и степень их раскрытия в документе. Степень раскрытия темы выражается положительным числом. Совокупность этих чисел составляет вектор, который будем называть *профилем документа*. Таким образом, каждый документ на сервере имеет соответствующий вектор, показывающий, насколько документ близок к той или иной теме.

Каждому пользователю сопоставляется подобный вектор, который будем называть *вектором интересов* или *профилем пользователя*. Если посетитель заходит на сервер в первый раз, то формируется его первоначальный вектор интересов по умолчанию. Один из способов задать его – создать из тех тем, которые наиболее популярны среди других посетителей.

Дальнейшие просмотры документов этим посетителем корректируют его профиль. Вариантов реализации также достаточно много, самый простой из них – сложение векторов текущего вектора интересов посетителя и профилей просмотренных документов.

Семантические сети и онтологии

Современные методы автоматической обработки данных, доступных в Интернете, как правило, основаны на частотном и лексическом анализе *текстового* содержимого [7, 8, 11, 14], которое прежде всего предназначено для восприятия человеком.

Однако более эффективным видится использование форматов описания, доступных для машинной обработки [3, 13].

Семантическая сеть – один из способов представления знаний в виде понятий и отношений. Это информационная модель предметной области, имеющая вид ориентированного графа, вершины которого соответствуют объектам предметной области, а дуги (рёбра) задают отношения между ними [1, 2]. Объектами могут быть понятия, события, свойства, процессы.

Семантические сети тесно связаны с понятием онтологии. Онтология – это попытка всеобъемлющей и детальной формализации некоторой области знаний с помощью концептуальной схемы. В информатике онтология должна иметь формат, который компьютер может легко обрабатывать.

На практике можно использовать семейство форматов RDF и OWL [16, 17]. Все объекты предметной области связываются специального вида утверждениями: “субъект-предикат-объект”, называемыми триплетами. Эти связи могут иметь самую разнообразную семантику – например, отражающую иерархию объектов. Однако, к сожалению, стандартный RDF-граф не является взвешенным, а для наших целей необходимо ввести веса связей, отражающие степень близости объектов и в дальнейшем работать с этими весами.

ПРОФИЛИРОВАНИЕ

Представление о профилировании

Профилирование или персонализация пользователя – это разумное ограничение предъявляемой пользователю информации с целью выделения более важного содержания для данного индивидуума. Задачей профилирования является правильный отбор пар «пользователь – набор отображаемых данных» путем отсеивания неинтересной пользователю информации. Решение этой задачи позволит потребителям услуг тратить меньше времени на просмотр информации и больше – на ее практическое применение.

Существует масса подходов к персонализации, но можно выделить два основных: персонализация пользователя путем изменения формы отображения данных и персонализация содержания – собственно, предъявляемых данных [18]. К первому подходу можно отнести, например, индивидуальную разметку страницы – сюда входит как расположение элементов друг относительно друга, так и цветовая гамма, темы оформления и так далее. Ко второму подходу относятся все способы выделения конкретных данных – более значимых для пользователя – по сравнению с остальными. В данной работе представлен именно метод персонализации Интернет-пользователей путем изменения набора отображаемых данных.

Связанность понятий. Представление о тематике

Поскольку понятия, затронутые в документе, могут быть крайне разнообразными, для их описания удобно использовать онтологии. Онтологии обладают рядом преимуществ: они моделируют знания, формальны, позволяют пользоваться всеми преимуществами семантической паутины. Для описания отношений между понятиями используем граф, в основе которого лежит RDF-граф онтологии, с одной важной модификацией. Как и в графе RDF, его узлы – это понятия, а ребра – связи (отношения) между ними. Отличие новой модели заключается в том, что каждой связи соответствуют два весовых коэффициента, каждый из которых означает вероятность не встретить одно понятие в контексте другого. Таким образом, это коэффициенты «различия», удаленности понятий друг от друга. Причины использования коэффициентов, определяющих степень различия между понятиями, вместо степени соответствия являются чисто техническими. Объяснение этому дано в конце параграфа.

Пример: связь между понятиями КОМПЬЮТЕР и ИНТЕРНЕТ имеет коэффициенты 0.6 и 0.15, означающие, что, когда речь идет о компьютерах, то в 40% (1 – 0.6 = 0.4) случаев в тексте встретится упоминание понятия Интернет, а когда документ касается глобальной паутины, в 85% случаев он будет иметь отношение к компьютерам.

Тематиками будем называть ёмкие понятия, то есть такие, которые тесно связаны с наибольшим количеством других понятий, что означает их важность в данной онтологии. Тематики мы будем также называть базовыми понятиями. Способ выделения тематик из всего набора понятий будет описан ниже.

Введем метрику, которая будет определять степень отношения понятия к той или иной тематике. Такой метрикой будет кратчайший путь в графе

до нее от рассматриваемого понятия. Эта метрика выбрана, потому что ее значение минимально для близких по смыслу или тесно связанных понятий и, наоборот, максимально для понятий, не имеющих друг к другу отношения, далеких друг от друга. Легко доказать, что наше построение удовлетворяет аксиомам метрики.

Именно для удобства обращения с данной метрикой мы использовали степень различия между понятиями вместо их связанности.

Определение тематики

Анализ показал, что тематики, то есть более общие понятия, могут быть выделены среди всех понятий по следующим критериям:

- понятие имеет большое количество входящих связей в RDF-графе. Это значит, что данное понятие связано с большим количеством других понятий, значит, оно достаточно распространенное, и, скорее всего, достаточно общее. Иначе, если бы оно было конкретным, дочерним понятием у более общего понятия, то встречалось бы наравне с другими конкретными, дочерними понятиями, и это было бы видно из анализа множества документов обучающей выборки. Необходимо принимать в расчет именно входящие связи, потому что большое число исходящих связей означает лишь то, что в данной онтологии понятие описано очень подробно, однако это ничего не говорит о его общности. Под указанный критерий подходят такие понятия, как, например, автомобиль – он составлен из множества деталей, потому в графе все детали будут иметь связи с этим понятием. И слово «автомобиль», скорее всего, будет часто упоминаться в документах рядом со смежными понятиями;
- понятие имеет множество потомков в дереве, образованном только связями «родитель-потомок». Под данный критерий попадают понятия, которые лежат в корне разветвленной иерархии. Самые первые объекты, порождающие иерархию, могут иметь мало входящих связей в графе, и таким образом не попадать под первый критерий. Тем не менее, находясь в корне иерархии, они самым естественным образом подходят на роль тематик – это самые общие понятия, которые конкретизуются своими потомками.

Чтобы выделить тематику, предлагается ввести систему набора «очков» для понятия по каждому критерию. Если понятие хорошо подходит под критерий – оно набирает большее количество очков. И если в сумме по критериям количество очков велико – понятие считается тематикой.

Вполне вероятно, что существуют и другие критерии выделения тематик. Однако балловая система хороша своей универсальностью, ее легко расширить новыми критериями путем простого добавления дополнительной проверки.

Определение значимости документа

Задача определения значимости документа для конкретного пользователя является важнейшей в данной работе, поскольку, решая ее, мы решаем задачу профилирования посетителя веб-сайта – выделяем документы, которые могут представлять для него наибольший интерес.

Для этого используем следующую модель: построим векторное пространство, осями которого являются тематики.

Векторы в этом пространстве – это:

- профили пользователей. Значение по каждой оси соответствует интересу пользователя по соответствующей теме;
- профили понятий и документов. Значение по каждой оси соответствует релевантности понятия/документа соответствующей тематике.

Степень соответствия документа интересам пользователя определяется корреляцией профилей пользователя и документа. Неплохой оценкой этой корреляции является скалярное произведение представляющих их векторов.

В данном случае пространство действительно должно быть векторным, то есть должны выполняться аксиомы векторного пространства, поскольку операции сложения векторов и умножения на скаляр необходимы при пересчете профилей пользователя и документа (см. ниже).

Определение профилей документов

Профиль документа строится на основе профилей входящих в него понятий. Как было сказано ранее, каждому понятию после некоторого предварительного анализа выборки документов ставится в соответствие вектор, характеризующий близость понятия к разным тематикам. Затем строим вектор документа. Самый простой способ построения профиля документа – сложить векторы входящих в него понятий. В результате получим вектор, каждая координата которого характеризует близость документа к той или иной тематике. Это ни что иное, как профиль документа по определению. Этот метод можно улучшить, учитывая особенности строения документа.

Например, если какое-то понятие встречается чаще остальных, можно увеличить его вклад в сумму векторов, умножая его на специальный коэффициент.

Определение профилей пользователей

При принятии решения, предъявлять ли пользователю тот или иной документ, необходимо учитывать:

- то, к каким тематикам пользователь *обычно* проявляет наибольший интерес, то есть его постоянные интересы;
- то, что интересно посетителю в данный момент. Это может быть тема, как имеющая отношение к кругу его постоянных интересов, так и не касающаяся их. Эта тематика должна быть представлена в наборе отображаемых документов наиболее широко.

Как было сказано ранее, профиль пользователя – это вектор, каждая координата которого – число – указывает величину интереса пользователя к какой-то конкретной тематике. Этот вектор будем называть базовым профилем пользователя $P_b = p_{b1}...p_{bn}$, он отражает интересы в долговременном масштабе – профессиональной направленности, хобби, увлечении.

Для отбора отображаемых документов используется так называемый оперативный профиль пользователя, который вычисляется покомпонентно, т.е. по каждой тематике в отдельности с учетом соответствующей составляющей базового профиля и количества документов, прочитанных посетителем по данной тематике за определенный отрезок времени до текущего момента. Оперативный профиль определяется следующим образом: при количестве просмотров (документов по заданной тематике) V(t) меньше определенного порогового значения V_{thr}, они не учитываются, чтобы отбросить ложные срабатывания. Если количество просмотров за заданный временной интервал не превышает значения C_{raise}, характеризующего повышенный интерес пользователя к тематике, значение составляющей оперативного профиля равно соответствующей компоненте базового профиля. Если же количество просмотров за заданное время превышает значение C_{raise}, то к оперативному профилю применяется так называемый повышающий коэффициент.

Более формально это можно записать так:

- $p_{oi} = 0$, при $V(t) < V_{thr}$
- $p_{oi} = p_{bi}$, при $V(t)/t <= C_{raise}$ или $t <= t_c$
- $p_{oi} = k_{raise} * p_{bi}$, при $V(t)/t > C_{raise}$ и $t > t_c$

Примечание: на самом деле в V(t) учитывается степень соответствия просмотренного документа рассматриваемой тематике, но для простоты на данном этапе считаем, что она соответствует количеству документов, в той или иной степени касающихся данной тематики, которые просмотрел пользователь.

Базовый профиль вычисляется следующим образом:

- при первом посещении сайта пользователю назначается начальный профиль со средними значениями интереса ко всем популярным среди других посетителей темам;
- в дальнейшем вектор корректируется: при просмотре очередного документа вектор его профиля умножается на некоторый коэффициент и добавляется к профилю пользователя, после чего производится нормировка последнего на 1.

Однако нужно учитывать некоторые особые ситуации и правильно их обрабатывать. Например, нужно уметь решать проблему «накрутки» профиля пользователя путем постоянного показа одного и того же документа. Профиль может довольно быстро и сильно отклониться в одну тематическую область. По-видимому, изменения должны происходить плавно и с некоторым затуханием. Этого можно добиться путем введения специальных коэффициентов ослабления.

Затухание интереса к теме

В ситуации, когда человек вдруг перестает интересоваться какой-либо темой, необходимо постепенно уменьшать коэффициент интереса к ней, и в какой-то момент вовсе переставать показывать соответствующие теме документы.

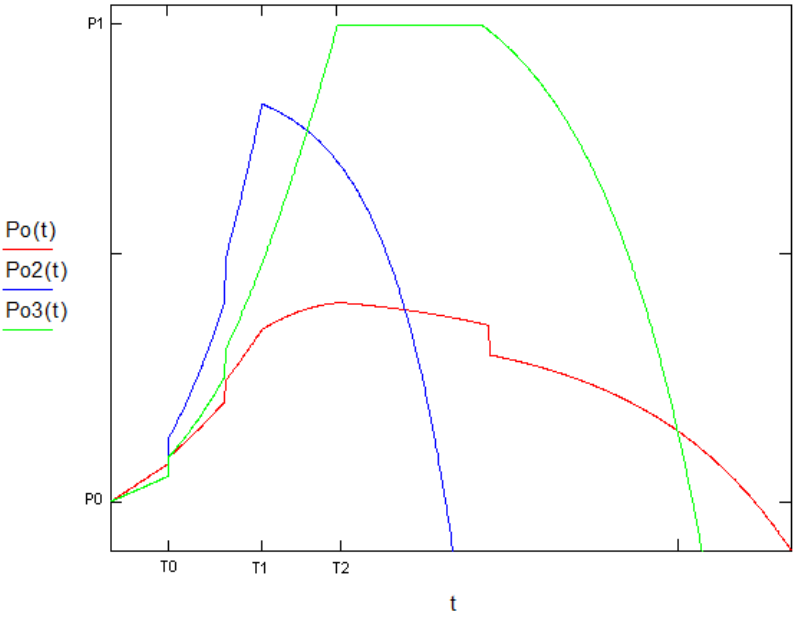
Предлагается уменьшать коэффициент за счет роста интереса к другим темам, то есть производить пересчет в момент просмотра новых документов. Таким образом, если человек не смотрит вообще никаких документов в течение пары месяцев, его профиль остается без изменений. Если же он смотрит другие документы, то интерес к старым темам постепенно замещается новыми.

Для начала можно ввести простой коэффициент, на который будем умножать текущее значение интереса. Например, при каждом новом просмотре умножать старые значения на 0.95.

Однако особым образом необходимо обрабатывать неожиданный временный всплеск интереса (описанный выше вариант, когда человек вдруг на пару дней стал остро интересоваться одной темой, и потом резко теряет

к ней интерес). Такие всплески нужно обрабатывать отдельно, чтобы снизить их влияние на постоянный профиль.

Ниже приведены графики зависимости интереса пользователя к какой-то теме от времени, при разном поведении пользователя (разная частота и количество просмотренных им документов). Графики были получены в программе Mathcad при моделировании поведения пользователя и его профилирования с помощью подхода, описанного в данной работе.



Зеленый график показывает рост профиля интереса, в случае если пользователь постоянно смотрит документы на заданную тематику, со средней частотой и длительностью, и его интерес не спадает. Однако, чтобы не допустить бесконечный рост одной темы, мы устанавливаем верхний порог для профиля. Как видно из зеленого графика, рост продолжается вплоть до времени t2 и максимального значения профиля P1, после чего график становится горизонтальным – достигается максимально возможное значение профиля.

Красный график – ситуация, когда пользователь смотрит документы на определенную тему вплоть до времени t2, а потом вдруг прекращает смот-

реть. В таком случае, за счет просмотра других документов уровень интереса по этой теме падает – как видно на красном графике, после времени t2 график стремится вниз.

Синий график показывает, как ведет себя профиль при неожиданном кратковременном всплеске интереса пользователя. Если в течение короткого времени с t0 до t1 посетитель просмотрел много документов по определенной теме, профиль растет резко. Но как только интерес к этой теме исчезает, профиль так же резко уменьшается. Как видно, синий график почти вертикально уходит вниз после времени t1.

Определение профилей понятий

Алгоритм подсчета профилей понятий

Определение профилей понятий является ресурсоемкой задачей, однако ее достаточно выполнить только один раз.

Модуль семантического разбора текста (*MCPT*) на входе получает текст, онтологию и словарь, на выходе – набор понятий из онтологии в той последовательности, в которой они встречаются в тексте. Модуль является сторонним [5], его разработка не является частью данной работы. В каждой из выходных последовательностей MCPT, полученных из «калибровочного» набора документов, производится подсчет пар понятий, попадающих в определенное «окно», то есть понятий, отстоящих друг от друга на расстоянии, не превышающем определенного значения – размера «окна», измеренного количеством понятий в выходной последовательности, стоящих между двумя рассматриваемыми. Данные заносятся в таблицу, строки и столбцы которой – понятия из онтологии. Пересечения – количество попаданий пары в «окно». Значения меньше определенного порога считаются случайными и отбрасываются [7].

Есть гипотеза, что значения в таблице коррелируют с RDF-связями между парами понятий. Вкратце ее обоснование можно свести к следующему: если связь между понятиями существует, она должна быть отражена в онтологии и присутствовать в документах. Следовательно, она будет обнаружена, и вероятность употребления пары понятий в одном контексте логично ляжет на RDF-связь между ними.

Присвоим каждой связи вес – значение из таблицы. Во взвешенном RDF-графе находим кратчайшие пути от каждого понятия до каждой из тематик. Для этого существуют стандартные алгоритмы: Дейкстры – для путей от одной вершины графа до всех остальных, Флойда–Уоршелла – для

путей между всеми вершинами графа, и другие. Расстояния в графе от понятия до каждой из тематик сохраняются в таблице, ряды которой – понятия, столбцы – тематики. Таким образом, у каждого понятия имеется профиль – вектор, описывающий, насколько данное понятие близко каждой из имеющихся тематик. На основе профилей упомянутых в документе понятий строится его собственный профиль.

Пример.

Пусть дан следующий текст:

«Математика – это наука, исторически основанная на решении задач о количественных и пространственных соотношениях реального мира путём идеализации необходимых для этого свойств объектов и формализации этих задач.

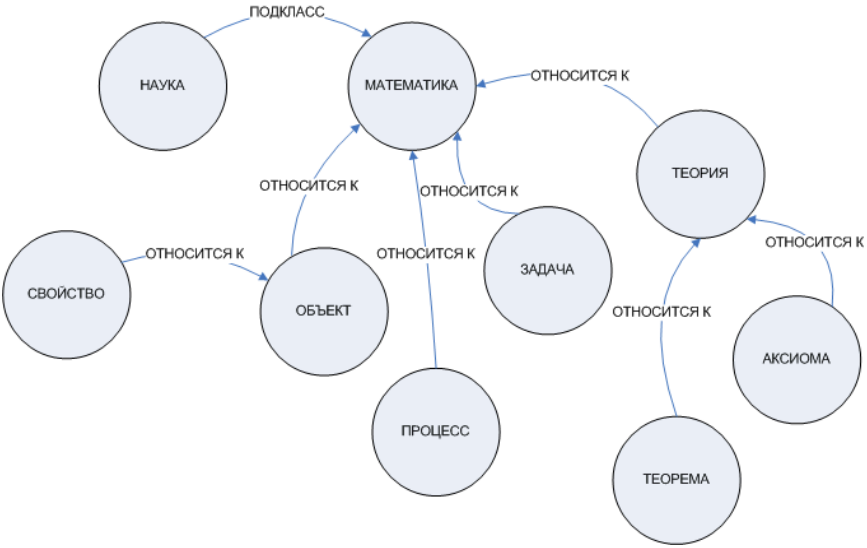
Обычно идеализированные свойства исследуемых объектов и процессов формулируются в виде аксиом, затем по строгим правилам логического вывода из них выводятся другие истинные свойства (теоремы). Эта теория в совокупности образует математическую модель исследуемого объекта. Т.о. первоначально исходя из пространственных и количественных соотношений, математика получает более абстрактные соотношения, изучение которых также является предметом современной математики».

Для анализа данного документа рассмотрим часть онтологии, описывающей математику. Онтология не очень подробная и включает только самые общие понятия, для показательного примера.

Применив компонент семантического разбора текста и используя в качестве словаря понятия из онтологии, мы выделяем из этого текстового фрагмента основные понятия:

«**Математика** – это **наука**, исторически основанная на решении **задач** о количественных и пространственных соотношениях реального мира путём идеализации необходимых для этого **свойств объектов** и формализации этих **задач**.

Обычно идеализированные **свойства** исследуемых **объектов** и **процессов** формулируются в виде **аксиом**, затем по строгим правилам логического вывода из них выводятся другие истинные свойства (**теоремы**). Эта **теория** в совокупности образует **математическую** модель исследуемого **объекта**. Т.о. первоначально исходя из пространственных и количественных соотношений, **математика** получает более абстрактные соотношения, изучение которых также является предметом современной **математики**».



Подсчитаем, сколько раз понятия попадали в одно окно вместе. Например, для данного текста возьмем окно равным 5 словам (при этом игнорируются так называемые «стоп-слова»).

...Обычно **[** идеализированные **свойства** исследуемых **объектов]** и **процессов** формулируются в виде **аксиом**, затем...

В данное окно вместе попали понятия **свойство** и **объект**.

Сдвигаем окно на одно слово влево:

...Обычно идеализированные **[свойства** исследуемых **объектов** и **процессов]** формулируются в виде **аксиом**, затем...

На этот раз в окно попали понятия свойство, объект, процесс.

Таким образом сканируем весь документ до конца. Заполним таблицу, в строках и столбцах которой – выделенные понятия. В ячейках – количество попаданий в одно окно.

	наука	мате-матика	зада-ча	тео-рия	объект	свой-ство	теорема	аксиома	про-цесс
наука		1	1						
математика	1			1	1				
задача	1				1	2			
теория		1					1		
объект		1	1			2		1	
свойство			2		2				1
теорема				1					
аксиома					1				
процесс						1			

Оценка сложности алгоритма

Самым затратным с точки зрения вычислительной сложности является алгоритм поиска кратчайших путей от каждого понятия до каждой из тематик в RDF-графе. Эта задача равносильна поиску кратчайших путей от каждой тематики до каждого понятия, то есть от некоторых вершин графа до всех остальных.

Наиболее подходящим для данной задачи из общеизвестных алгоритмов выглядит алгоритм Дейкстры. Он находит кратчайшее расстояние от одной из вершин графа до всех остальных за время $O(n^2 + m)$, где n – количество вершин, а m – количество ребер в графе. Оценив количество ребер в графе, как n^2 – одна связь от каждой вершины к каждой, получим оценку сложности $O(n^2)$. Эту операцию необходимо провести p раз, где p – количе-

ство тематик. Тогда сложность будет $O(pn^2)$. Если предположить, что $p=O(n^{1/2})$, сложность алгоритма будет составлять $O(n^2n^{1/2}) = O(n^{5/2})$. Это значение не слишком велико для онтологий с количеством понятий до 100000: для такой онтологии потребуется порядка 10^{13} операций, что составит около часа работы компьютера с производительностью порядка 10^{10} теоретических операций в секунду, например, настольного компьютера на базе Intel Core 2 Quad [15]. Поскольку данная задача требует вычисления лишь один раз, дополнительных оптимизаций не требуется.

ЗАКЛЮЧЕНИЕ

В работе рассмотрен метод профилирования пользователей Интернет-порталов с использованием семантических сетей для повышения релевантности набора предлагаемых к прочтению документов интересам конкретного посетителя. Метод заключается в определении тематик, представляющих интерес для пользователя; оценивается близость того или иного документа к этим тематикам. Профиль пользователя динамически меняется в зависимости от набора прочтенных им документов. Кроме того, кратковременные и постоянные интересы пользователя веб-сайта обрабатываются различным образом. Теоретическая модель проиллюстрирована наглядными примерами.

СПИСОК ЛИТЕРАТУРЫ

1. **Berners-Lee T., Hendler J., Lassila O.** The Semantic Web // Scientific Am. – 2001. – N 3. – P. 34–43.
2. **Berners-Lee T., Shadbolt N., Hall W** The Semantic Web Revisited // IEEE Intelligent Systems. – 2006. – N 3. – P. 96–101.
3. **Ильин Н.И. и др.** Технологии извлечения знаний из текста // Открытые системы. – 2006. – № 6 – С. 14–17.
4. **Демин А. В.** Логико-вероятностный метод извлечения знаний и его применение в задачах прогнозирования и управления: Автореф. дис... канд. физ.-мат. наук: 05.13.11. – Новосибирск, 2008. – 18 с.
5. **Сидорова Е. А.** Методы и программные средства для анализа документов на основе модели предметной области: Автореф. дис... канд. физ.-мат. наук: 05.13.11. –Новосибирск, 2006. – 19 с.
6. **Baker D., Mccallum A.** Distributional Clustering of Words for Text Classification // Proc. of the 21st Annual Internat. ACM SIGIR Conf. on Research and

Development in Information Retrieval, Categorisation. – Melbourne, 1998. – P. 96–103.
7. **Некрестьянов И.С.** Тематико-ориентированные методы информационного поиска Дис. ... канд. физ.-мат. наук: 05.13.11. – СПб., 2000
8. **Некрестьянов И.С., Добрынин В.Ю., Клюев В.В.** Оценка тематического подобия текстовых документов // Тр. второй всероссийской научной конф. «Электронные библиотеки». – Протвино, 2000. – С. 204–210.
9. **Aumueller D., Rahm E.** Caravela: Semantic Content Management with Automatic Information Integration and Categorization // European Semantic Web Conf. – Innsbruck, 2007. – P. 729–738.
10. **Извозчикова В.В. и др.** Один из подходов к поиску информации на основе семантических сетей // Перспективные информационные технологии и интеллектуальные системы. – 2004. – № 3. – С. 23–27.
11. **Ferreira A., Atkinson J.** Intelligent Search Agents Using Web-Driven Natural-Language Explanatory Dialogs // Computer. – 2005. – Vol. 38, N 10. – P. 44–52.
12. **Capra R., Perez-Quinones M.** Using Web Search Engines to Find and Refind Information // Computer. – 2005. – Vol. 38, N 10. – P. 36–42.
13. **Ding L. et al.** Search on the Semantic Web // Computer. – 2005. – Vol. 38, N 10. – P. 62–69.
14. **Семенов Ю.А.** Современные поисковые системы [Электронный ресурс]. – Телекоммуникационные технологии. / ГНЦ ИТЭФ, 2003 – Гл. 4.5.14. – Режим доступа: https://shu.ru/~avr/tt/retr4514.shtml.htm, свободный. – Загл. с экрана.
15. **Единицы** измерения производительности процессоров Intel® в соответствии с экспортными требованиями [Электронный ресурс]. – Intel Corporation, 2009. – Режим доступа: http://www.intel.com/support/ru/processors/sb/CS-023143.htm#3, свободный. – Загл. с экрана.
16. **Klyne G. et al.** Resource Description Framework (RDF): Concepts and Abstract Syntax. W3C Recommendation. [Электронный ресурс] / W3C, 2004. – Режим доступа: www.w3.org/TR/rdf-concepts/, свободный. – Загл. с экрана.
17. **Patel-Schneider P.F. et al.** OWL Web Ontology Language Semantics and Abstract Syntax. W3C Recommendation. [Электронный ресурс] / W3C, 2004. – Режим доступа: http://www.w3.org/TR/owl-semantics/, свободный. – Загл. с экрана.
18. **Скуратов А.К., Ефремов С.В.** Персонификация и персонализация как основа современных порталов // Телематика'2003: Тр. X Всероссийской научно-методической конф. – 2003. –Т. 1. – С. 183–185
19. **Buchwalter C., Ryan M., Martin D.** The state of online advertising: data covering 4th Q 2000. – TR Adrelevance, 2001.
20. **Salton G., McGill M. J.** Introduction to modern Information Retrieval. – McGraw-Hill, 1983.

А.В. Будний, А.Д. Русанов

МЕТОД ПРОФИЛИРОВАНИЯ ПОЛЬЗОВАТЕЛЕЙ ИНТЕРНЕТ-ПОРТАЛОВ СРЕДСТВАМИ SEMANTIC WEB

Рукопись поступила в редакцию 29.10.09
Редактор Т. М. Бульонкова
Рецензент Ю. А. Загорулько

Подписано в печать 11.11.09
Формат бумаги 60 × 84 1/16 Объем 1.1 уч.-изд.л., 1.25 п.л.
Тираж 60 экз.

Центр оперативной печати «Оригинал 2»
г.Бердск, ул. О. Кошевого, 6, оф. 2, тел. (383-41) 2-12-42