Методика учета интересов пользователя при работе в сети Internet на основе его профиля и ассоциативных связей

© М. М. Шарнин © А. В. Петров © И. П. Кузнецов
Институт Проблем Информатики РАН,
Москва
mc@keywen.com apetrov@keywen.com igor-kuz@mtu-net.ru

Аннотация

В статье рассматривается методика, обеспечивающая выдачу полезной информации пользователю Интернет при его обращении к поисковым системам. Имеется в виду упорядочение ответов в соответствии с интересами пользователя, а также выдача рекламы, которая может заинтересовать пользователя. Для этого вводится понятие «профиль пользователя», состоящий из лексических единиц, отражающих интересы пользователя. Выявление таких единиц осуществляется путем лексического анализа текстов запросов, а также текстов сайтов, которые заинтересовали пользователя при получении ответа на запрос. Такие единицы образуют признаки, характеризующие пользователя. Помимо сказанного, методика использует дерево категорий, где в каждой категории имеются ссылки на сайты (и на рекламу). Из текстовых компонент этих сайтов также извлекаются лексические единицы, которые образуют признаки, характеризующие данную категорию. Эти признаки (называемые первичными) дополняются новыми признаками (называемые вторичными), полученными за счет ассоциативных связей категории, которые выявляются методами дистрибутивной семантики.

Методика, обеспечивающая выдачу полезной информации (в том числе, рекламы), заключается в сопоставлении признаков, входящих в профиль пользователя, и признаков из дерева категорий с выбором адекватных категорий, которые определяют все последующие выдачи. Использование профилей пользователей позволяет получать хорошие результаты даже на коротких поисковых запросах.

1 Введение

Одно из направлений повышения качества поисковых систем (Интернет-систем), работающих в сети Интернет, связано с использованием знаний о характеристиках пользователей. Такими характеристиками могут быть пол, возраст, интересы пользователей в виде признаков и др. Обладая более полной информацией о пользователе, поисковые системы могут обеспечить более “персонализированные” выдачи, а именно:

- Показывать пользователю интересующий его контент (статьи, музыку, фильмы, книги и т.д.);
- Подстраивать поисковые выдачи под пользователя, например, поисковая система может показывать более интересные пользователю страницы на более высоких позициях;
- Оптимизировать выдачу рекламы, показывая пользователю только те страницы и картинки, которые ему интересны.

Как правило, данные об особенностях и характеристиках конкретного пользователей отсутствуют в явном виде в сети Интернет. В тоже время Интернет-системы располагают большим набором данных о действиях пользователя при работе с компьютером, на основании которых можно с достаточной точностью восстановить недостающую информацию. Данные о действиях включают в себя:

- Историю визитов пользователя по различным интернет-страницам;
- Историю поисковых запросов пользователя;
- Историю кликов по гиперссылкам и баннерам;
- Историю покупок в интернет-магазинах;
- Историю просмотра видеороликов;
- Историю прослушивания аудиофайлов.

В данной статье при определении характеристик пользователя будут рассматриваться только два типа действий: история визитов и история поисковых запросов. Будет предложен метод, позволяющий находить характеристики

пользователей автоматически путем лингвистического анализа поисковых запросов и посещаемых пользователем веб-страниц.

2 Существующие подходы

В целом, работы, так или иначе связанные с определением интересов пользователей, можно разделить на две большие группы.

- Методы не использующие лексику.

Такие методы, как правило, базируются на сравнении действий пользователя при работе в сети Интернет с действиями других пользователей с нахождением “близких пользователей”. Методы, использующие близость действий, называются «коллаборативной фильтрацией» и широко используются для подбора товаров в интернет-магазинах [1], для выдачи рекомендаций (например, при выборе музыкальных произведений) [2] и др.

- Методы, использующие лексику.

В этом случае анализируется тексты поисковых запросов пользователя, а также тексты интернет-страниц, посещаемых пользователем. Далее на выделенных текстах используются алгоритмы автоматического рубрицирования, позволяющие соотнести тексты к тому или классу [3, 4], который и будет характеризовать пользователя. При этом используются:

o Алгоритмы использующие косинусную меру между текстами;

o Метод опорных векторов.

В международном проекте Large Scale Hierarchical Text classification (LSHTC) Pascal Challenge [1], проводившемся осенью 2009 года, наилучший алгоритм по доле правильно классифицированных документов заключался в делении документов на кластеры (классы) и определении близости документов к центрам классов. При этом в качестве функции близости был использован косинус угла между векторами признаков документа и кластера. Следует отметить, что косинус угла между векторами – это по существу линейное правило объединения признаков документа.

3 Представление интересов пользователя

Для решения задачи, заключающейся в выявлении интересов пользователя, был выбран подход, основанный на автоматическом анализе истории пользовательских запросов и посещаемых им веб-сайтов. Выбирались текстовые документы, которые были использованы как обучающая выборка. На основе последней путем автоматической рубрикации выделяются признаки (лексические единицы, термины), характеризующие данного пользователя.

Интересы пользователя можно представить в виде набора пар <термин, вес>. Например, покупательница может иметь следующие интересы: <платье, 70> <косметика, 40> <сумочка, 20>, а интересы спортивного болельщика могут быть представлены так: <спорт, 30> <футбол, 60> <воллейбол, 20>. Здесь цифрами представлен вес термина, выражающего конкретный интерес, в некоторой шкале. Отметим, что интересы конкретного пользователя могут быть также представлены в виде точки (или вектора) в многомерном векторном пространстве признаков. Образуется вектор пользовательских интересов. Такое представление позволяет ввести меру близости между интересами различных пользователей (по близости точек в пространстве терминов), и в результате автоматически определять группы пользователей с близкими интересами.

Для классификации интересов пользователей использовалось дерево категорий (взятое из Интернет), где каждая категория соответствует определенной тематике, выражается с помощью лексических единиц, терминов или понятий (например, ФУТБОЛ, МИНИФУТБОЛ, АМЕРИКАНСКИЙ ФУТБОЛ) и имеет ссылки на соответствующие поясняющие тексты (сайты, статьи). Эти тексты используются как обучающая выборка, из которой путем автоматической рубрикации выделяются признаки (лексические единицы, термины), связанные с данной категорией. Они представляются в виде троек <признак, категория, вес>, где вес – это некоторое число, которое характеризует значимость признака для определения категории. Будем называть такие признаки первичными. Количество таких троек может быть расширено путем включения терминов, связанных ассоциативными связями с термином-категорией. При этом поиск связей осуществляется по множеству текстов, где упоминается данный термин, например, ФУТБОЛ. Будем называть такие термины вторичными. Появляются тройки <вторичный_термин, категория, вес>, где вес определяет силу ассоциативной связи.

С каждой категорией связывается соответствующий контент или реклама. Поэтому одной из важных задач является соотнесение пользователя к определенными категориями упомянутого дерева. Такое соотнесение сводится к сравнению признаков, характеризующих пользователя, и признаков, связанных с категориями. По их близости выделяются категории, к которым и соотносится тот или иной пользователь. Ему выдается соответствующий контент или реклама. При сравнении признаков учитывается не только их совпадение, но и наличие ассоциативных связей (за счет вторичных признаков).

Отметим, что зная вектор пользовательских интересов, можно сделать предположение о половой принадлежности и возрасте пользователя. Например, футбол интересует больше мужчин, косметика – женщин. Современная музыка чаще интересует молодежь, а ретро чаще интересует пожилых людей. Вектора пользовательских интересов, дополненные информацией о категориях пользователя, мы называем профилем пользователя.

4 Построение векторов пользовательских интересов

Как уже говорилось, на основе истории запросов и визитов пользователя по различным Интернет-страницам осуществляется выбор текстов (из запросов и соответствующих страниц в Интернете). Путем анализа и статистической обработки текстов строится вектор в многомерном векторном пространстве, компонентами которого являются частоты различных терминов, присутствующих в этих текстах. Это и есть вектор пользовательских интересов.

Множество признаков каждой категории, взятой из дерева категорий (с учетом вторичных признаков), также можно представить в виде многомерного вектора в том же векторном пространстве, в котором были представлены пользовательские интересы. Веса соответствующих признаков должны быть тщательно рассчитаны, чтобы обеспечить наиболее точную классификацию текстовых документов и историй пользователей.

Главная идея лучших алгоритмов вышеописанного международного проекта LSHTC состояла в том, что каждому первичному признаку (термину) в каждой категории присваивается свой вес. При этом первичные признаки пополняются новыми признаками, найденными за счет ассоциативных связей (см. выше). Далее осуществляется поиск этих признаков в документах, которые наиболее часто посещал пользователь. Подсчитываются веса признаков, входящих в документ. На основе них находится наилучшая категория для документа, которая присваивается пользователю.

5 Расчет оптимальных весов вторичных признаков с категорией

Задачу автоматической классификации документов (их соотнесения к категориям), связанных с пользователем, можно свести к задаче коллективного решения автоматических экспертов, каждый из которых будет реагировать только на присутствие в документе одного конкретного признака (термина), связанного с категорией. Таким образом количество экспертов будет совпадать с количеством таких признаков и каждый эксперт будет отвечать только за один конкретный признак. Если признак присутствует в документе, то эксперт принимает решение о принадлежности документа к заданной категории. Вероятность правильного решения такого эксперта совпадает с вероятностью присутствия указанного признака в документах заданной категории.

Главным результатом работы Nitzan и Paroush (1982), а также Shapley и Grofman (1984) [http://www.socsci.uci.edu/] было утверждение, что если вероятность правильного решения каждого эксперта известна, то линейное правило объединения их решений является оптимальным, и максимальная вероятность правильного коллективного решения достигается когда весовые коэффициенты (акции) экспертов рассчитываются по формуле

$$Wi = \log(Pi/1-Pi),$$

где Pi – это вероятность правильного решения эксперта с номером «i».

6 Расчет вероятности соотнесения признака с категорией

В простейшем случае в качестве вероятности соотнесения признака (термина) к категории может служить частота употребления признака в текстах, связанных с данной категорией, деленная на частоту признака в целом во всех анализируемых текстах. Данный способ близок к широко распространенной мере TF-IDF [5], которая учитывает не только частоту термина в текстовых документах, связанных с категорией, но еще и частоту документов с данным термином, что уменьшает вес у общеупотребимых и малозначимых слов, таких как междометия, предлоги и т.д. Более точный расчет вероятности вхождения термина в контекст рубрики должен учитывать ошибку расчета вероятности.

7 Расчет весов первичных признаков, связанных с категорией

В качестве обучающей выборки для формирования начального варианта дерева категорий может быть использована одна из многочисленных интернет директорий, содержащих ссылки на веб-сайты, например, Google, Yahoo, Yandex.

В данной работе в качестве эталонного каталога русскоязычных сайтов использовался Яндекс.Каталог. Яндекс.Каталог представляет из себя древовидный набор рубрик, каждая из которых содержит в себе краткое описание, подрубрики и ссылки на веб-страницы. При этом одна веб-страница может описывать несколько рубрика. При этом, если веб-страница принадлежит рубрике (имеется на нее ссылка), то она так же принадлежит всем рубрикам верхнего уровня.

Например, рубрика ФУТБОЛ (верхний уровень – СПОРТ) состоит из подрубрик:

РОССИЙСКИЙ ФУТБОЛ
МИРОВОЙ ФУТБОЛ
МИНИ ФУТБОЛ . . .,

И ссылается на сайты

ФУТБОЛ НА ПООРТАЛЕ «ЧЕМПИОНАТ.СОМ»
РОССИЙСКИЙ ФУТБОЛЬНЫЙ СОЮЗ, . . .

Сайты имеют поясняющие тексты и картинки, по которым идет обращение к сайтам.

Для сбора текстовой информации с web-страниц использовалась программа на языке python и библиотека nltk<вставить ссылку>. С помощью стандартных функций библиотеки nltk текст web-страниц был очищен от html-разметки и разбит на цепочки слов. Из таких цепочек были выделены словосочетания, играющие роль признаков, длиной не более 50 символов. Для каждого словосочетания было посчитано количество вхождений в документ. В итоге для каждой web-страницы из эталонного каталога был сформирован файл, состоящий из троек <словосочетание, рубрика, количество вхождений>.

На следующем шаге при помощи программы обработки данных map-reduce [6] для каждой пары (<словосочетание>, <рубрика>) было посчитано общее количество вхождений словосочетаний в каждую из рубрик. Итоговый файл использовался для присвоения рубрикам их признаков. Он имел ту же структуру и формат, что и документы для каждой из web-страниц.

По этим данным были рассчитаны вероятности вхождения словосочетаний в документы, связанные с рубриками. По этим вероятностям при помощи специальной формулы (см. ниже) рассчитывались веса признаков, связанных с рубриками.

Вес вторичного признака (словосочетания, термина), связанного с рубрикой, рассчитывается через вероятность вхождения признака в контекст основных признаков рубрики. Под контекстом здесь понимается множество всех документов и их компонент, содержащих основные признаки рубрики.

Извлечение ассоциативных связей из Интернет текстов базируется на дистрибутивной гипотезе, утверждающей, что семантически близкие (или связанные) признаки (термины,словосочетания) имеют похожий контекст и, наоборот, при похожем контексте признаки семантически близки.

Порождение вторичных признаков с их весами предполагает использование различных методов, включающих:

- методы выявления из Интернет текстов определенных предметных областей;

- методы выявления из текстов значимых словосочетаний, терминов и их ранжирования;

- методы выявления и ранжирования ассоциативных связей между значимыми словосочетаниями и терминами.

Методы предполагают предварительное обучение на текстах, в том числе, взятых из различных Интернет-ресурсов.

Обработка больших массивов текстов, постоянно пополняемых в среде Интернет, позволяет собирать необходимые статистические данные для формирования достаточно полной картины предметной области, представленной в виде набора ассоциативных связей. Возможность проводить машинное обучение на большом количестве примеров придает системе определенную гибкость и улучшает результаты.

8 Согласование признаков, рассчитанных разными методами

Вес признака, связанного с категорией, может быть рассчитан различными методами, например, по вероятности вхождения признака (термина, словосочетания) в документы, на которые ссылается категория. Следующий возможный метод – это расчет веса через вероятность нахождения признака в контексте основных признаков категории. Под контекстом здесь понимается множество всех документов и их компонент, содержащих первичные признаки категории.

Веса признаков, рассчитанных разными методами, должны по возможности совпадать. Для этого был предложен следующий метод согласования весов.

Допустим с помощью двух различных методов мы нашли две группы признаков T1 и T2, связанных с некоторой категорией. Например, T1 – это признаки, входящие в тестовую выборку, на которую ссылается категория (это основные термины и их меньше), а T2 – это признаки, связанные с категориями ассоциативными связями (т.е. это вторичные признаки) и выделенные по Интернет-текстам (их больше).

Веса рассчитываются независимо для T1 и T2 двумя различными алгоритмами. Но так как множество T1 входит в T2, то веса для T1 будут рассчитаны обоими алгоритмами и мы подбираем коэффициент для весов из T2, при котором эти веса максимально приближаются к весам из T1. В дальнейшем этот коэффициент используется для подсчета весов вторичных признаков.

9 Определение интересов пользователя по истории его посещений и запросов

Ранее было показано, каким образом для каждой из посещенных пользователем web-страниц можно выявить их тематику – соотнесение к категориям. Это может служить основой для построения профиля пользователя, отражающие его интересы. Такой профиль можно представить в виде множества пар <категория, вес>. В качестве веса можно, например, взять отношение количества web-страниц данной категории «C» к общему количеству web-страниц, посещенных пользователем.

Предположим, что интересы пользователя неизменны. Тогда можно считать, что количество web-страниц категории «C» в общем объеме страниц подчиняется биномиальному распределению. Пусть «N» – общее количество документов, просмотренных пользователем; а «n» - количество документов категории «C». В этом

случае вес будет вычисляться по формуле W= n/N. При этом доверительный интервал (с уровнем доверия 95%) можно вычислить по формуле:

Wmin = W - 1.96 *SQRT(W*(1-W)/N).

Wmax = W + 1.96 *SQRT(G2*(1-G2)/D2)

Например, если n = 2, а N=4, то в этом случае Wmin = 0.01, а Wmax = 0.99, то есть ничего конкретного про значение W сказать нельзя. Если же n=100, а N=200, то Wmin = 0.43, а Wmax = 0.56 — в этом случае можно с уверенностью говорить, что обнаружена категория и что это не «шум».

Для того, чтобы исключить из сферы интересов пользователя «шумы» (т.е. web-страницы, на которые пользователь зашел случайно), можно включать в профиль пользователя только те интересы, для которых минимальное значение доверительного интервала больше некоторого порога.

10 Заключение

В данной статье был рассмотрен метод, позволяющий превратить историю посещений пользователем сайтов и историю поисковых запросов в историю его интересов и намерений. Была разработана программа, которая успешно определяла интересы пользователя, в том числе, по коротким текстам, не содержащим терминов их обучающей выборки (построенной на основе дерева категорий). Отметим, что типовые алгоритмы, например описанные в проекте LSHTC, плохо работают на таких текстах.

В дальнейшем планируется разработать алгоритм, который позволит по этой информации получить профиль пользователя, учитывающий множество факторов, в том числе, его пол, возраст, интересы, намерения, регион проживания, уровень доходов, семейное положение и другую полезную информацию. Работа выполнена при поддержке гранта РФФИ 13-07-00272.

Литература

[1] Greg Linden, Brent Smith, Jeremy York “Amazon.com recommendations. Item-to-item collaborative filtering”. EEE Internet Computing, Los Alamitos, CA USA, 2003 http://www.cs.umd.edu/~samir/498/Amazon-Recommendations.pdf

[2] Brian McFee, Luke Barrington, Gert Lanckriet “Learning Similarity from Collaborative Filters”, ISMIR 2010, http://cosmal.ucsd.edu/cal/pubs/ISMIR2010_learnCF.pdf

[3] Агеев М. С. «Методы автоматической рубрикации текстов, основанные на машинном обучении и знаниях экспертов», диссертация: 05.13.11, Москва, 2004

[4] Абрамов В. Е. «Автоматическое рубрицирование и реферирование текстовой информации : в том числе на иностранных языках», диссертация: 05.25.05, Москва, 2008.

[5] Salton, G. and McGill, M. J. 1983 Introduction to modern information retrieval. McGraw-Hill, ISBN 0-07-054484-0.

[6] J Dean, S Ghemawat, «MapReduce: simplified data processing on large clusters», Communications of the ACM, 2008 - dl.acm.org.

[7] E.Baharad, J.Goldberger, M.Koppel и S.Nitzan, “Beyond Condorcet: Optimal Aggregation Rules Using Voting Records”, CESifo München, 2011.

[8] A. Lenci, “Distributional semantics in linguistic and cognitive research”, Rivista di Linguistica, 1, 2008, pp.1-30.

[9] M.Baroni, A.Lenci, “Distributional Memory: A General Framework for Corpus-Based Semantics”, Computational Linguistics. V.36, Issue 4, 2010, pp. 673-721.

[10] Peter Turney, “A uniform approach to analogies, synonyms, antonyms and associations”, Proceedings of COLING, Manchester, 2008, pp. 905–912.

[11] Peter Turney, “Mining the web for synonyms: PMI-IR versus LSA on TOEFL”, Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001), Freiburg, Germany. September 3–7, 2001. pp. 491–502.

[12] Michael Charnine, “Keywen Automated Writing Tools”, Booktango, USA, 2013, ISBN 978-1-46892-205-9.

Method for identification of Internet users interests based on associations and users profiles

M.Charnine, A.Petrov, I.Kuznetsov

This article describes the method for providing useful information for user of Internet search engines. The method allows to organize search results and related advertisements according to user’s interests. Here we introduce the concept of "user profile", consisting of keywords and terms, reflecting user interests. The discovering of such keywords and terms is done by parsing user queries and visited websites. The proposed method uses a tree of categories linked to related advertisements and websites. From these websites we retrieve primary keywords characterizing categories. The primary keywords are extended with associated secondary keywords which were obtained by the methods of distributive semantics. We determine relevant categories, useful information and related advertisements by comparing keywords from user's profile and keywords of the categories. The use of distributional semantics methods allows us to obtain good results even on short search queries.