УДК 004.738.52

СИСТЕМА ФОРМИРОВАНИЯ ИНФОРМАЦИОННОГО БАЗИСА КОРПОРАТИВНЫХ СИСТЕМ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ

Наукенова Т.А.
научный руководитель канд. техн. наук Зеленков П.В.
Сибирский федеральный университет

Аннотация

Основным преимуществом распределенного информационного поиска перед простым информационным поиском является использование данных, полученных от многих поисковых систем, использующих разные алгоритмы для поиска и ранжирования данных. Такой подход позволяет существенно улучшить и результаты поиска, и приводит к сокращению времени, затрачиваемого на поиск необходимой информации. Для увеличения качества результатов поиска и выдвижения на первые позиции результатов, наиболее релевантных для конкретного пользователя используется профиль пользователя.

Ключевые слова: Информационный поиск, метапоиск, персонифицированный поиск, профиль пользователя, ранжирование результатов поиска, слияние данных.

Введение

Информационный поиск давно стал частью нашей повседневной жизни. Приоритетной задачей в настоящее время является построение механизма эффективного ранжирования результатов поиска, отсеивания информации, заведомо не релевантной для осуществляющего поиск с целью повышения эффективности поиска. Построение такого механизма невозможно без персонализации поиска. В данной статье описывается работа по построению профиля пользователя и его использования для перестроения результатов не персонифицированного метапоиска. Эффективные алгоритмы слияния и ранжирования данных, полученных от многих поисковых систем совместно с алгоритмами персонификации, основанными на предпочтениях и учитывающими интересы конкретного пользователя позволят получить максимально эффективные результаты и существенно сократить время, затрачиваемое пользователем на поиск необходимой информации.

Описание системы

В моделируемой информационной системе предприятия, часть структуры которой представлена на рис.1, профили пользователя создаются администратором системы. Профиль содержит следующую информацию: фамилия, имя, отчество сотрудника; возраст; пол; департамент; должность; дополнительная информация. Пользователи связаны между собой иерархическими отношениями (руководитель департамента-подчиненный), пользователи имеют возможность устанавливать неформальные отношения между профилями (друзья). ИС предоставляет возможность создавать группы по интересам (работа над проектом, группы по интересам). Предприятие имеет свои вики-ресурсы, сообщества. В ИС предусмотрена возможность хранения документов с функцией комментирования. Рейтинг интересов пользователя представлен набором терминов, которые близко связаны с пользователем.

Формирование списка терминов

Если пользователь является автором документа, или участвует в обсуждении документа или статьи вики-ресурса (внутри группы или непосредственно на форуме) теги, используемые в этом документе, попадают в интересы пользователя. Попытка реализовать персонификацию информационного поиска с использованием кругов общения пользователя рассматривается в работе [1,2]. Для получения информации о поведении пользователя вне внутренней сети, необходимо анализировать историю просмотров интернет-страниц. Сбор данных происходит при помощи дополнения к

информационной системе предприятия [4]. В механизме этого дополнения предусматривается фильтрация анализируемых данных (исключается анализ паролей пользователя). Для извлечения терминов применяется алгоритм, описанный в [6], использующий комбинацию лингвистической и статистической информации, чтобы оценить каждый термин. Термины – кандидаты находятся с помощью ряда языковых моделей, затем им присваиваются весовые значения, в зависимости от частоты появления термина или его подтерминов в тексте.

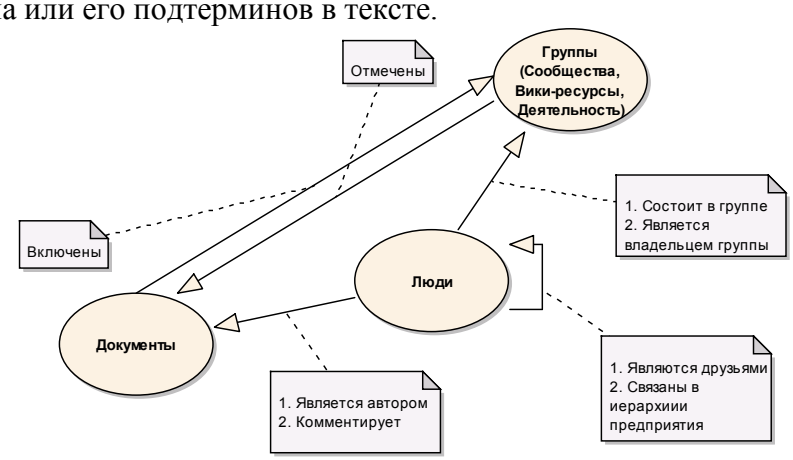

Рис.1

Оценивание значимости терминов

После того, как список терминов получен, вычисляется вес каждого из них тремя способами. I способ - присвоение веса по частоте встречаемости. Необходимо определить вектор частоты, который содержит количество появлений термина t_i во всех источниках данных, как показано в (1).

$$\vec{F}_{t_i} = \begin{bmatrix} f_{title_{t_i}} \\ f_{mdest_{t_i}} \\ f_{text_{t_i}} \\ f_{mkeyw_{t_i}} \\ f_{terms_{t_i}} \\ f_{nphrases_{t_i}} \end{bmatrix} \quad (1)$$

Где $\vec{F}$ - вектор частоты появлений термина t_i , f_{title_t} - количество появлений термина t_i во всех названиях за всю историю просмотра страниц, f_{mdest} - количество появлений t_i во всех метаданных внутри тегов <*meta name="description"*>, f_{text} - количество появлений t_i в текстах страниц, f_{mkeyw} - содержимое тегов <*meta name="keywords"*>, f_{terms} - количество появлений t_i, в словах, определенных с помощью алгоритма, описанного в работе [6] как ключевые термины страницы, $f_{nphrases}$ - количество появлений t_i во всех именных фразах на странице. Вес термина будет определяться по следующей формуле с использованием вычисленного вектора:

$$\omega_{TF}(t_i) = \vec{F}_{t_i} \cdot \vec{\alpha} \quad (2)$$

Где значение вектора $\vec{\alpha}$: 0 – игнорировать поле, 1 –включить поле, 1/Ni, где Ni – итоговое количество появлений термина в этом поле i. Таким образом больше веса дается терминам в более коротком поле (например, в поле ключевых слов).

II вариант, рассматриваемый в описываемой модели, это TF-IDF (TermFrequency, InverseDocumentFrequency) взвешивание. При этом подходе, слова, появляющиеся во многих документах, спускаются вниз в рейтинге значимых по инвертированной частоте появления:

$$\omega_{TFIDF(t_i)} = \frac{1}{\log(F_{t_i})} \times \omega_{TF(t_i)} \tag{3}$$

Где $\omega_{TF}(t_i)$ - вес термина по частоте встречаемости (из формулы (2)).

Последний применяемый метод взвешивания – модификация метода BM25, данная в работе [5]:

$$\omega_{pBM25}(t_1) = \log \frac{(r_{t_i} + 0.5)(N - n_{t_i} + 0.5)}{(n_{t_i} + 0.5)(R - r_{t_i} + 0.5)} \tag{4}$$

Где N представляет количество документов в веб (по оценкам GoogleN-Gram, 220 680 773), n_{t_i} - количество документов в разделе, включающем термин (оценивается с использованием GoogleN-Gram), R–количество документов в истории просмотров данного пользователя, и r_{t_i} - количество документов из истории просмотра пользователя, которые содержат искомый термин.

Формирование поискового профиля

Персонифицированный профиль для поиска генерируется «на лету», как только пользователь авторизуется в системе. Списки значимых терминов профиля пользователя, полученных из локальной сети предприятия, объединяются со списками терминов, полученными из истории просмотра веб-страниц. Если термины встречаются и в первом и во втором списках, то веса этих терминов суммируются. Далее происходит перестроение полученного списка результатов поиска в соответствие со списком терминов и их весов.

Модели слияния данных

Целью работы является скомбинировать результаты поиска N поисковых систем, работающих с одним и тем же запросом, в итоговый ранжированный список документов. Для достижения этой цели предлагается ранжировать документы по их вероятности появления согласно разработанной модели. Рассматриваемый подход использует корректировки Байеса для получения максимальной апостериорной оценки для вероятности появления каждого документа согласно модели.

Модель Нетерпеливого Читателя

Чтобы воспользоваться порядком ранжирования документов каждой отдельной системы вводится идея «нетерпеливый читатель» [3]. Нетерпеливый читатель – это гипотетический пользователь, который задал системе запрос, в ответ на который система выдала s_i, ранжированный список документов от системы S_i. Получив эти результаты, пользователь должен просматривать их, изучать в том порядке, в котором они выданы системой. Предполагается, что пользователь утомляется в процессе чтения, и в какой-то момент принимает решение остановиться. Наиболее вероятно, что пользователь просмотрит документ, находящийся на 1 позиции, нежели на 50.

Будем моделировать вероятность достижения читателем документа, используя предположение, что интерес пользователя иссякнет довольно быстро. Таким образом моделируется вероятность достижения пользователем документа с рангом r в списке документов с экспоненциальным распределением:

$$f(r \mid \lambda) = \lambda e^{-\lambda r} \tag{5}$$

Где $\lambda \geq 0$ – параметр, отвечающий за то, как быстро пользователь утомится.

Уравнение 5 позволяет нам вычислить вероятность того, что читатель достигнет заданной позиции r в ранжированном списке документов .

Слияние данных с использованием корректировки Байеса

В предыдущей части работы были очерчены два источника информации, используемые для получения оценки θ_j . MLE основано на числе систем, возвративших документ d_j. Но мы также получаем некоторую информацию о документе в связи с его положением в ранжированном списке $s_1 \ldots s_N$. Чтобы учесть эту информацию можно использовать корректировку Байеса для получения максимальной апостериорной оценки θ_j . Оценивание с использованием Баесовских корректировок дает возможность успешного сочетания двух видов информации - мощность документа среди N систем и средняя вероятность из модели нетерпеливого читателя. Ранние наработки о Байесовском оценивании содержатся в работе [1].

Взвешенная смешанная модель

Часто случается, что мы располагаем информацией об относительной эффективности каждой из входных систем $S_1 \ldots S_N$. Обычно в таком случае оценивают данные, пришедшие от предположительно наиболее эффективной системы в процессе слияния. В нашем случае наиболее релевантными будут считаться документы, полученные из внутренней сети предприятия. Потому мы добавим в построенную модель коэффициент, указывающий, что внутренние документы должны иметь ранг больший, нежели документы, пришедшие из внешней сети.

Заключение и последующие работы

В данной работе рассмотрена модель профиля пользователя для персонификации информационного поиска в рамках информационной системы предприятия с учетом истории поведения пользователя в сети интернет. Рассмотрен механизм получения информации о поведении пользователя в сети, ее обработки и получения необходимых данных. Полученные данные объединяются в один профиль, который дополняется в режиме реального времени, что позволяет поисковой системе учитывать все происходящие в поведении пользователя изменения. Персонифицированный профиль пользователя используется для перестроения результатов поиска с учетом всех полученных из профиля пользователя персональных данных. Результаты, полученные при выполнении данной работы, имеют существенное значение для развития моделей и методов поиска и обработки информации при управлении сложными информационно-управляющими системами производственного назначения.

Использовнная литература

1. 2 William M. Bolstad. Introduction to Bayesian Statistics. Wiley Interscience, New York, NY, 2007.
2. 3 D. Carmel, N. Zwerdling, I. Guy, S. Ofek-Koifman, N. Har’el, I. Ronen, E. Uziel, S. Yogev, S. Chernov. “Personalized Social Search Based on the User’s Social Network”, 2009
3. 6 M. Efron, Generative Model-Based MetaSearch for Data Fusion in Information Retrieval, 2009
4. 4 N. Matthijs, F. Radinski. “Personalizing Web Search using Long Term Browsing History”, 2011
5. 5 J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In Proc. of SIGIR, pages 449-456, 2005
6. 1 A. T. A. Thuy Vu and M. Zhang. Term extraction through unithood and termhood unification. InProc. of Int'l Joint Conf on Natural Language Proc., 2008.