УДК 004.89

Ю. О.Трегубова, А.С. Вороной
Донецкий национальный технический университет, г.Донецк
кафедра систем искусственного интеллекта

АЛГОРИТМЫ ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ ТЕМАТИЧЕСКОГО ПОИСКА В ИНТЕРНЕТ

Аннотация

Трегубова Ю. О., Вороной С.М. Алгоритмы повышения эффективности тематического поиска в Интернет*. Рассмотрен метод тематического поиска, основанный на повышении релевантности результатов поиска по ключевым словам путем применения процедур фильтрации и классификации. Рассмотрены модификации алгоритмов Байеса и разделяющих гиперплоскостей, обеспечивающие низкую сложность вычислений*

Ключевые слова: *тематический поиск, фильтрация, классификация, метод Байеса, разделяющая гиперплоскость, низкая сложность вычислений* **.**

Постановка проблемы. В области информационного поиска отдельно выделяется задача тематического поиска, то есть целенаправленного поиска документов, относящихся с высокой степенью релевантности к определенной теме, заявленной пользователем. Широко применяемые в Интернет машины поиска по ключевым словам, малоэффективны с точки зрения поиска тематической информации из-за большого уровня шума (ссылок на нерелевантные документы), ограниченных возможностей языков запросов и формы представления результатов поиска [1]. В этих условиях разработка метода тематического поиска в Web, повышающего качество поиска по сравнению с традиционными методами в условиях долговременности информационной потребности пользователя и динамичности пространства поиска, представляется актуальной.

Цель статьи – провести анализ метода повышения эффективности тематического поиска и предложить модификации алгоритмов тематически-ориентированной классификации текстовых документов

Описание метода тематического поиска. Информационная потребность пользователя представляется в виде пары $\{q, D\}$, где q – запрос по ключевым словам, использующийся для первичного отбора документов из Web, $D = \{D+, D-\}$ – обучающая выборка, описывающая тему, интересующую пользователя. Данная обучающая выборка содержит примеры релевантных теме документов ($D+$) и нерелевантных документов ($D-$).

Общий вид технологии тематического поиска на основе предлагаемого метода представлен на рис. 1.

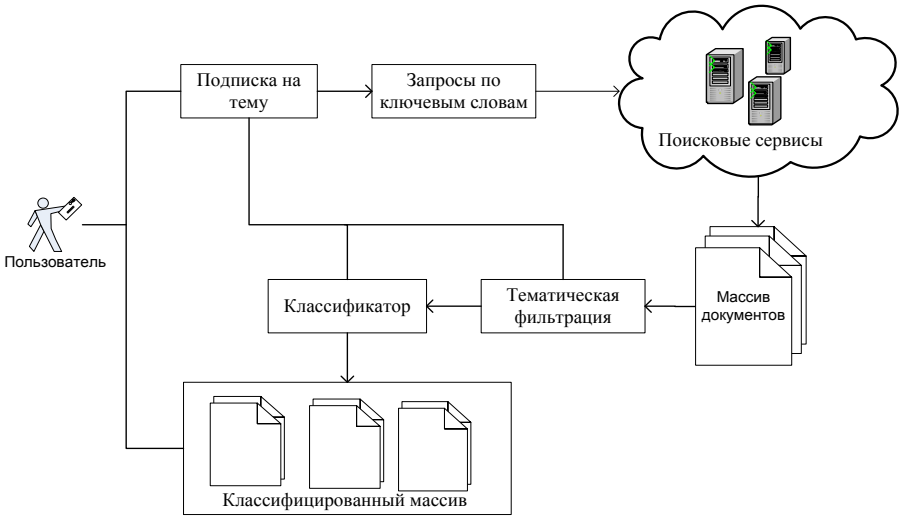

Рисунок 1 - Основные этапы тематического поиска

Процесс поиска реализуется в два этапа:

1. Отбор документов из Web, соответствующих запросу по ключевым словам *q* с помощью глобальных систем поиска по ключевым словам. Данный этап позволяет с одной стороны обеспечить высокую полноту поиска, а с другой – существенно сократить объем обрабатываемой на следующем этапе информации.

2. Уточнение результатов поиска с помощью классификатора, обученного на предоставленной пользователем обучающей выборке *D*. Этот этап позволяет обеспечить высокую точность результатов поиска.

Для реализации классификации результатов поиска пользователю необходимо в обучающей выборке множество релевантных документов *D*+ разбить на подмножества, описывающие интересующие пользователя подтемы. В этом случае обучающая выборка будет представлять собой множество $D=\{D_1+, D_2+, D_3+,\ldots, D_n+, D-\}$, где D_i+ - обучающая выборка i-ой подтемы, *n* – общее количество подтем.

Таким образом, в этом случае классификатор будет решать две задачи: задачу тематической фильтрации (*бинарной классификации*) и задачу разбиения множества релевантных теме документов на подтемы (задачу классификации с большим количеством классов в обучающей выборке).

Рассмотрим более детально второй этап метода - задачу уточнения результатов поиска.

Формальная постановка задачи классификации текстов выглядит следующим образом.

Предполагается, что алгоритм классификации работает на некотором множестве документов $D=\{d_i\}$

Все множество документов разбивается на непересекающиеся подмножества классов

$$C = \{C_i\}, \bigcup_{d \in C_i}^{\forall i} d = D, C_i \cap C_j = 0 (i \neq j)$$

Задачей классификации является определение класса, к которому относится данный документ.

Предварительная обработка документов. Задачей этапа предварительной обработки документов является выделение признаков документа и сопоставления им весов. В простейшем случае набором признаков документа будет содержащийся в нем набор лексем, а в качестве веса используется количество вхождений лексемы в документ. При обработке Web-страниц необходимо предусмотреть автоматическое определение кодировки, поскольку явно она указывается не во всех документах. Для лексем, входящих в заголовки, названия, ключевые слова, текст ссылки и т.п. производится увеличение веса.

Алгоритмы классификации для этапа уточнения результатов поиска. Для обеспечения масштабируемости и приемлемой вычислительной сложности можно использовать в качестве базовых алгоритмы Байеса и разделяющих гиперплоскостей с линейной (O(N)) сложностью обучения.

Модификация алгоритма Байеса. Метод Байеса это простой классификатор, основанный на вероятностной модели, имеющей сильное предположение независимости компонент вектора признаков [2,3]. Обычно это допущение не соответствует действительности и потому одно из названий метода - Naıve Bayes (Наивный Байес). Простой байесовский классификатор относит объект X к классу C_i тогда и только тогда, когда выполняется условие: $P(C_i|X)>P(C_j|X)$, где $P(C_i|X)$ – апостериорная вероятность принадлежности объекта X к классу C_i, $P(C_j|X)$ – апостериорная вероятность принадлежности объекта X к произвольному классу C_j, отличному от C_i. Т.е. апостериорная вероятность принадлежности объекта к классу C_i больше апостериорной вероятности принадлежности объекта к любому другому классу.

Алгоритм Байеса в настоящее время оценивается как сравнительно низкокачественный алгоритм. Основными причинами являются проблемы, связанные с принципом независимости признаков и некорректной оценкой априорной вероятности в случае существенно неравномощных обучающих выборок.

Правило определения класса для документа в алгоритме Байеса можно представить следующим образом:

$$C(d) = \arg\max_{C} [\log(p(C) + \sum_{w \in d} f_w \log p_{Cw}],$$

где f_w - количество вхождений признака w в документ,

$$p_{Cw} = p(w \mid C)$$

Для борьбы с низким качеством, используется парадигма класса-дополнения, то есть вместо вероятности принадлежности лексемы классу оценивается вероятность принадлежности лексемы классу-дополнению C' (следует учесть, что p(w|C) ~ 1-p(w|C')).Используя принцип сглаживания параметров по Лапласу, получаем следующее правило:

$$C(d) = \arg\max_C [\log(p(C) - \sum_{w \in d} f_w \log(\frac{\overline{N}_{Cw} + 1}{\overline{N}_C + |V|})]$$

где $\overline{N}_{Cw}$ - количество вхождений признака во все классы кроме данного, $\overline{N}_C$ - общее количество вхождений всех признаков в класс-дополнение, $|V|$ - размерность словаря признаков.

Данная эвристика работает только в том случае, если количество классов $|C| >> 2$.

Алгоритм решения задачи бинарной классификации. Для задачи бинарной классификации рассмотренные модификации не позволяют приблизить метод Байеса по качеству к лучшим показателям т.к. использование классов-дополнений не дает никаких изменений, поэтому для данного случая предлагается использовать алгоритм, основанный на линейном дискриминанте Фишера. В основе алгоритма лежит поиск в многомерном признаковом пространстве такого направления *w*, чтобы средние значения проекции на него объектов обучающей выборки из классов максимально различались.

Мерой разделения спроецированных точек служит разность средних значений выборки. Желательно, чтобы проекции на прямой были хорошо разделены и не очень перемешаны. Проекцией произвольного вектора X на направление *w* является отношение $(wX^{t})/|w|$). В качестве меры различий проекций классов на *w* используется функционал – индекс Фишера[4,5]:

$$\Phi(\mathbf{w}) = \frac{[\tilde{\mathbf{X}}_{w1}(\mathbf{w}) - \tilde{\mathbf{X}}_{w2}(\mathbf{w})]^2}{\tilde{d}_1(\mathbf{w}) + \tilde{d}_2(\mathbf{w})},$$

$$\tilde{X}_{wi}(\mathbf{w}) = \frac{1}{m_i} \sum_{s_j \in \tilde{S}_t \cap K_i} \frac{(\mathbf{w}\mathbf{x}_j^t)}{|\mathbf{w}|},$$

$$\tilde{d}_i(\mathbf{w}) = \frac{1}{m_i} \sum_{s_j \in \tilde{S}_t \cap K_i} [\frac{(\mathbf{w}\mathbf{x}_j^t)}{|\mathbf{w}|} - \tilde{X}_{wi}]^2 ,$$

где $\tilde{X}_{wi}(W)$ - среднее значение проекции векторов, описывающих объекты из класса K_i, $i \in \{1,2\}$;

$\tilde{d}_i(W)$ - выборочная дисперсия проекций векторов, описывающих объекты из класса.

Линейный дискриминант Фишера определяется как линейная разделяющая функция $w_t x$, максимизирующая функционал Ф(W). Для более полного разделения классов в алгоритме строится несколько направлений соответствующих дискриминанту Фишера. Вдоль одного направления можно эффективно разделить часть обучающих экземпляров. Точки отсечения для положительных и отрицательных экземпляров вдоль каждого направления запоминаются при обучении алгоритма.

Схема обучения алгоритма представлена на рис.1.

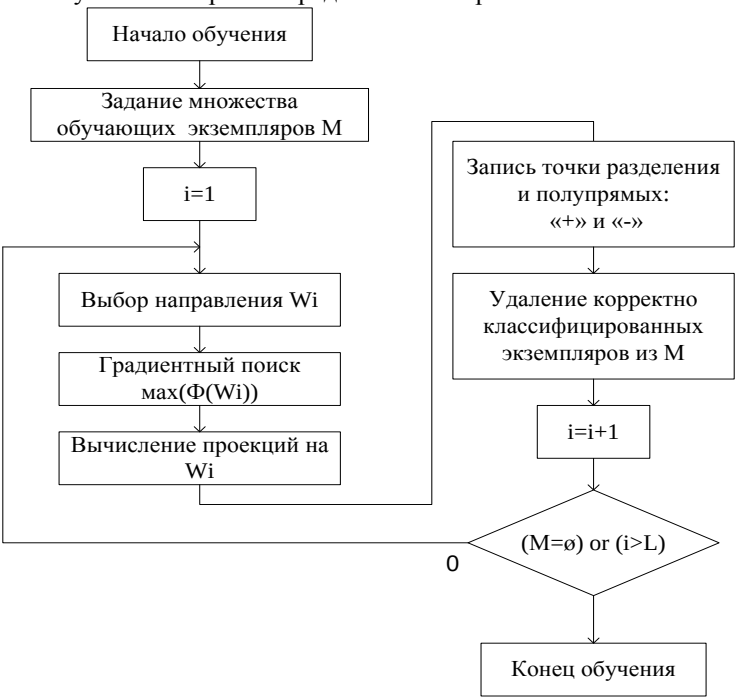

Рисунок 2 - Блок схема обучения алгоритма бинарной классификации

Классификация экземпляра производится по следующему алгоритму:

цикл i = 1... L (количество направлений)

Анализируем *i*-ое направление:

- *если* документ находится на полупрямой положительных или отрицательных документов, выдаем соответствующую метку и выходим из цикла
- *если* данное направление последнее, определяем метку экземпляра с помощью точки оптимального разделения классов

конец цикла

Предлагаемые направления совершенствования алгоритмов. В дальнейшем предполагается усовершенствование подготовки документов к применению методов машинного обучения путем выбора более эффективной формы оценки весов признаков документа, учитывающей ссылки. Проведение экспериментов по усовершенствованию байесовского классификатора документов для тем с большим количеством неравномощных подтем. Разработка критериев оценки качества поиска с применением предложенных алгоритмов.

Выводы

Проведен анализ существующих подходов к повышению релевантности тематического поиска путем классификации результатов выдачи поисковых систем. Разработаны алгоритмы фильтрации и классификации с малой вычислительной сложностью. Классификация результатов поиска позволяет существенно сократить время поиска нужной информации. Таким образом, введение дополнительной классификации на получаемые пользователем документы позволяет повысить удобство использования поисковой системы и позволяет быстрее ориентироваться в полученных результатах.

Список литературы

1. Ландэ Д.В. Поиск знаний в Internet. - М.:Диалектика, 2005. - 272 с.
2. Агеев М. С. Методы автоматической рубрикации текстов, основанные на машинном обучении и знаниях экспертов. Дис. канд. физ-мат. наук: 05.13.11. Московский гос. унив. - Москва, 2005.
3. Sebastiani F. Machine Learning in Automated Text Categorization. URL: http://nmis.isti.cnr.it/sebastiani/Publications/ACMCS02.pdf
4. R. Fisher. The use of multiple measurements in taxonomic problems. *Eugen.*, 7:179-188, 1936.
5. Максаков А.В. Масштабируемые алгоритмы классификации текстов// Труды 12-й конференции "Математические методы распознавания образов" (ММРО-12), Москва, 2005.