Об одном методе периодического тематического поиска информации в Web

© А.В. Максаков

Московский государственный университет им. М.В. Ломоносова
bruzz@yandex.ru

Аннотация

В статье описывается метод периодического тематического поиска, основанный на композиции метода поиска по ключевым словам и тематической фильтрации с использованием классификаторов текстов. Рассматриваются различные алгоритмы классификации, с точки зрения эффективности их применения при решении рассматриваемой задачи.

1 Введение

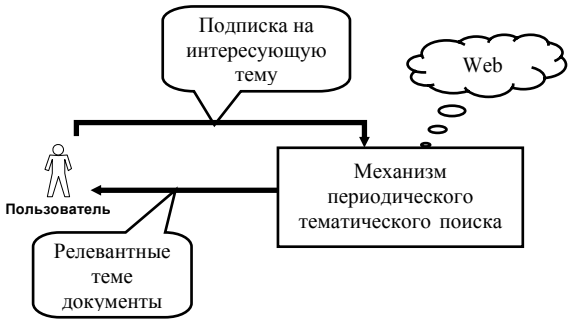
Бурное развитие сетевых технологий, в том числе и сети Интернет приводит к значительному увеличению доступных информационных ресурсов и объемов передаваемой информации. Зачастую это разнородная, слабо структурированная и избыточная информация, обладающая высокой динамикой обновления. Необходимость эффективного использования этого колоссального и динамично изменяющегося объема информации обуславливает актуальность и значимость исследований в области информационного поиска.

В области информационного поиска отдельно выделяется задача тематического поиска, то есть такого поиска, в рамках которого информационная потребность пользователя выражается некоторой определенной темой. Одним из возможных вариантов доставки релевантной информации является периодическая доставка, по аналогии с подпиской на тематические издания. Для обеспечения такого рода доставки информации из Web необходимо решить задачу периодического тематического поиска, то есть такого поиска, при котором множество новых документов, относящихся к заданной теме, предоставляется не сразу, а через определенные заранее промежутки времени. При этом следует отметить, что сервис периодической доставки релевантных документов имеет смысл только в том случае, если тематическая потребность остается актуальной и неизменной в течение большого промежутка времени.

2 Метод решения задачи периодического тематического поиска, основанный на использовании классификаторов

2.1 Описание и особенности задачи

С точки зрения пользователя процесс поиска будет выглядеть следующим образом:



Отличительными особенностями задачи периодического поиска в контексте данной работы является следующее:

- Высокая динамичность пространства поиска
- Информационная потребность пользователя сохраняется неизменной большой промежуток времени.
- Не требуется немедленное предоставление результатов поиска, однако существуют временные ограничения на время поиска.

Данные особенности позволяют ослабить требования, предъявляемые к методам периодического тематического поиска, по сравнению с традиционными методами информационного поиска. В частности, можно использовать методы, обладающие большей вычислительной сложностью поиска, поскольку не требуется немедленная выдача результатов. Условие статичности информационной потребности пользователя делает оправданным с практической точки зрения более подробное описание этой потребности.

2.2 Описание предложенного метода

В данной работе предлагается метод решения задачи периодического тематического поиска в Web, основанный на комбинации поиска по ключевым словам и тематической фильтрации с использованием классификаторов текстов, применяемой в ограниченных по объему коллекциях документов. Отбор документов при помощи поиска по ключевым словам позволяет на порядки сократить множество анализируемых классификатором документов, что приводит к уменьшению вычислительной сложности метода в целом по сравнению с методом тематической фильтрации и, как следствие, применимости полученного метода на больших объемах данных, характерных для Web.

Информационная потребность пользователя представляется в виде пары $\{q, D\}$, где

q – запрос по ключевым словам (запрос по КС), 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 ($D-$).

Процесс поиска разделяется на 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.

Ряд исследований [3,4] показали, что 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.

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).

2.3 Типовые показатели качества поиска

Традиционными показателями качества поиска являются *полнота* и *точность*. Определим эти показатели. Предположим D_c - множество документов, соответствующих заданной теме C, $s(q,M)$ - множество документов, найденных с помощью метода M. При этом тема описана некоторым запросом q.

Определение полноты тематического поиска.

Полнота (recall) тематического поиска определяется как соотношение количества документов, правильно отнесенных к теме к общему количеству документов, относящихся к данной теме:

$$P(s(q,M)) = \frac{|\, s(q,M) \cap D_c \,|}{|\, D_c \,|}$$

Определение точности тематического поиска.

Точность (precision) тематического поиска определяется как соотношение количества документов, правильно отнесенных к теме к общему количеству найденных документов:

$$R(s(q,M)) = \frac{|\, s(q,M) \cap D_c \,|}{|\, s(q,M) \,|}$$

Для того чтобы сравнивать качество различных методов информационного поиска, необходимо ввести интегральный показатель качества. В 1979 году Рийсбергеном была предложена такая мера (F-мера), имеющая в условиях равнозначности полноты и точности поиска вид [12]:

$$F1(s(q,M)) = \frac{2 \cdot P(s(q,M)) \cdot R(s(q,M))}{P(s(q,M)) + R(s(q,M))}$$

Следует отметить, что большой объем доступной информации в Web и ее изменчивость обуславливают невозможность точной оценки качества поиска. Как правило, для оценки качества поиска в Web используют оценки качества поиска по первым N документам из списка результатов поиска.

2.4 Оценка качества поиска

Выразим качество поиска с помощью предложенного метода (обозначим его $M_{гибр}$) через качество поиска по ключевым словам $M_{кс}$ и качество тематической фильтрации $M_{кл}$. Точность предложенного метода будет равна точности тематической фильтрации $M_{кл}$, используемой на втором этапе поиска. Полнота же будет

определяться произведением полноты отбора документов Web по ключевым словам *q'* на полноту тематической фильтрации:

$$P(M_{гибр}) = \frac{N_{\substack{релевантных \\ найденныз}}}{N_{отобранных}} = P(M_{кл}) \quad (2.4.1)$$

$$R(M_{гибр}) = \frac{N_{\substack{релевантных \\ найденных}}}{N_{релевантных}} =$$

$$\frac{N_{\substack{релевантных \\ отобранныз}}}{N_{релевантных}} \cdot \frac{N_{\substack{релевантных \\ найденныз}}}{N_{\substack{релевантных \\ отобранных}}}$$

или

$$R(M_{гибр}) = R(q', M_{КС}) \cdot R(M_{кл}) \qquad (2.4.2)$$

Поскольку качество поиска напрямую зависит от качества используемого классификатора, то алгоритм классификации должен обеспечивать высокую полноту и точность тематической фильтрации и классификации. Наличие ограничений на время обработки анализируемого множества документов обуславливает важность низкой вычислительной сложности классификации документов. Пользователь осуществляет обратную связь с системой периодического тематического поиска путем внесения изменений в обучающую выборку, что приводит к дообучению или полному переобучению классификатора. Это, в свою очередь, определяет важность низкой вычислительной сложности обучения классификатора.

2.5 Обоснование предложенного метода

В качестве альтернативы предложенному методу можно рассматривать *периодический поиск по ключевым словам* и *периодическую тематическую фильтрацию*.

Тематическая фильтрация с использованием классификаторов обладает более высокой вычислительной сложностью по сравнению с фильтрацией на основе запроса по ключевым словам: O(|V|) против O(|q|) на документ, где |V| - размерность словаря, используемого при классификации, |q| - количество термов в запросе по ключевым словам, при этом на практике |V| >> |q|. Такая высокая вычислительная сложность приводит к практической неприменимости тематической фильтрации на основе классификаторов в условиях, когда только для индексирования появившихся новых страниц необходимо иметь канал, способный обеспечить скачивание не менее чем 10 мегабайт текста в секунду [1].

Современные системы поиска по ключевым словам позволяют получить результаты за очень короткое время (обычно в пределах секунд и долей секунды). Однако с точки зрения качества поиска, у такого подхода есть недостатки. Если рассматривать три показателя: полнота поиска, точность поиска и время на составление запроса, то на практике можно достичь хороших результатов только для двух показателей [8].

Покажем, что при выполнении определенных условий предложенный метод будет превосходить по качеству поиска, выраженному мерой F1, метод поиска по ключевым словам.

Запрос по КС, результаты поиска по которому обладают наилучшим среди данного множества запросов Q показателем меры F1, будем называть F1-оптимальным запросом на этом множестве.

$$q = \arg\max_{q \in Q} F1(q)$$

Полноту поиска с помощью F1-оптимального запроса обозначим *R(q)*, точность – *P(q)*.

Определим следующие условия:

1. Применяемый классификатор превосходит по полноте F1-оптимальный запрос на множестве запросов Q, которые может предложить пользователь, т.е.

$$\exists \alpha > 1 : R(M_{кл}) \geq \alpha \cdot R(q) \qquad (2.5.1),$$

2. F1-оптимальный запрос обладает полнотой меньше единицы

$$R(q) < 1 \qquad (2.5.2)$$

3. Возможно подобрать запрос по КС q', уменьшающий ошибку, связанную с полнотой в произвольное количество раз

$$\exists q' : R(q') > R(q) \qquad (2.5.3)$$

Лемма: Предположим, что выполняются условия (2.5.1)-(2.5.3), причем

$$R(q') \geq R(q) + b \cdot (1 - R(q))$$

Тогда

$$\forall \alpha > 1 \quad \exists b < 1 : R(M_{гибр}) = R(q') \cdot R(M_{кл}) \geq R(q)$$

Доказательство:

Предположим, условие

$$R(M_{гибр}) = R(q') \cdot R(M_{кл}) \geq R(q)$$

выполняется. Получим оценку параметра *b*:

$$\alpha \cdot R(q) \cdot (R(q) + b \cdot (1 - R(q)) \geq R(q).$$

$$R(q) + b \cdot (1 - R(q) \geq 1/\alpha$$

$$b \geq \frac{1/\alpha - R(q)}{1 - R(q)}$$

Поскольку $\alpha > 1$, то $\frac{1/\alpha - R(q)}{1 - R(q)} < 1$. Что требовалось доказать.

Следствие 1. При выполнении указанных условий можно подобрать такой запрос по ключевым словам, что полнота гибридного подхода будет превосходить полноту поиска по F1-оптимальному запросу.

Множество Q может быть составлено двумя способами:

- Путем включения множества *ad-hoc* запросов, предложенных пользователями для описания заданной темы

- Путем получения запроса с помощью алгоритма *индукции правил*, обученного на представленной пользователем обучающей выборке.

Согласно [7,10,13] для запросов полученных с помощью алгоритмов индукции правил условие (2.5.1) выполняется на практике в большинстве случаев. При этом рассматриваемые алгоритмы превосходят алгоритмы индукции правил и по точности классификации. Результаты апробации, приведенные в разделе 4.2, подтверждают выполнение условия (2.5.1) и для множества представленных пользователем *ad-hoc* запросов. Таким образом, если используемый классификатор превосходит по точности F1-оптимальный запрос, качество результатов гибридного подхода, выраженное мерой F1 в среднем будет выше, чем качество поиска по ключевым словам, при условии подбора для отбора документов из Web запроса по КС, обладающего достаточно высокой полнотой.

С практической точки зрения преимущество предложенного метода состоит в том, что для получения высокого качества поиска необходимо составить запрос по ключевым словам обеспечивающий высокую полноту поиска, в отличие от инженерного подхода, в котором требуется составление запроса, обеспечивающего одновременно высокую полноту и точность.

Более подробно экспериментальное сравнение качества предложенного подхода и поиска по ключевым словам будет рассмотрено в разделе 4.2.

3 Решение задачи классификации

Исходя из (2.4.1) и (2.4.2) качество предложенного метода поиска напрямую зависит от качества используемого классификатора. Рассмотрим более подробно проблему решения задачи классификации, с учетом указанных в разделе 2.4 требований.

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.

3.1 Алгоритмы классификации, используемые на этапе уточнения результатов поиска

Основным выводом из нескольких независимых публикаций [13, 14] является преимущество одного из алгоритмов классификации – SVM[6] (метода опорных векторов) над другими алгоритмами.

Основным недостатком этого алгоритма остается относительно высокая вычислительная сложность обучения ($O(N^a)$ [2], где a > 1,7, N – количество документов в обучающей выборке).

В работе предложены два масштабируемых алгоритма классификации, обладающих линейной (O(N)) вычислительной сложностью обучения:

1. Модифицированный алгоритм Байеса, для решения задачи классификации с большим количеством классов в обучающей выборке.
2. Алгоритм построения нескольких разделяющих гиперплоскостей (*ModFisher*) для решения задачи бинарной классификации.

3.2 Модификация алгоритма Байеса

Экспериментальные исследования поведения алгоритма Байеса позволили обнаружить два систематических недостатка, сильно понижающих качество классификации:

- Предпочтение классификатором классов, содержащих большее количество примеров в обучающей выборке.
- Предпочтение классификатором классов, в которых содержится большее количество взаимно зависимых признаков (не выполняется предположение о независимости признаков).

Для борьбы с некорректным определением параметров, в случае неравномощных обучающих выборок, предлагается использовать парадигму класса-дополнения. В этом случае, вместо вероятности принадлежности признака классу оценивается вероятность принадлежности признака классу-дополнению C’ (следует учесть, что чем меньше вероятность принадлежности документа классу-дополнению p(d|C’), тем больше вероятность его принадлежности исходному классу p(d|C)) . Используя принцип сглаживания параметров по Лапласу, получаем следующее правило определения метки класса:

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

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.

Для частичной компенсации использования принципа независимости признаков, производится нормализация весов признаков

$$вес_{cw} = \frac{\log(\theta_{cw})}{\sum_{w \subset C} |\log(\theta_{cw})|}, \text{ где } \theta_{Cw} = \frac{\overline{N}_{Cw} + 1}{\overline{N}_C + |V|}$$

В результате при задаче классификации в случае наличия большого количества классов (|C| >> 2) предложенный алгоритм сравним по качеству классификации с алгоритмом SVM и существенно превосходит базовый алгоритм.

3.3 Метод построения нескольких разделяющих гиперплоскостей

Для задачи бинарной классификации внесенные модификации не позволяют приблизить метод Байеса по качеству к лучшим показателям (парадигма классов-дополнений не вносит никаких изменений), поэтому для данного случая предлагается использовать алгоритм с условным названием ModFisher. Идея алгоритма состоит в последовательном нахождении направлений (как правило, не более 3-4), соответствующих *дискриминанту Фишера*[2, 5], максимизирующему так называемый *индекс Фишера*

$$J(a) = \frac{(\frac{1}{|x|}\sum_{x\in X}(x,a) - \frac{1}{|y|}\sum_{y\in Y}(y,a))^2}{\frac{1}{|x|}\sum_{x\in X}(x,a)^2 - (\frac{1}{|x|}\sum_{x\in X}(x,a))^2 + \frac{1}{|y|}\sum_{y\in Y}(y,a)^2 - (\frac{1}{|y|}\sum_{y\in Y}(y,a))^2}$$

Вдоль такого направления можно эффективно разделить часть обучающих экземпляров. В дальнейшем возможно два сценария: либо спроецировать все экземпляры на полученное малоразмерное пространство и использовать внешний алгоритм, например C4.5 [11], либо использовать точки отсечения для положительных и отрицательных экземпляров вдоль каждого направления. В ходе экспериментов второй вариант показал более точные результаты.

Схема обучения алгоритма выглядит следующим образом:

1. Методом градиентного спуска находим локальный максимум *J(a)*.
2. Проецируем все обучающие экземпляры на полученное направление и запоминаем точку оптимального разделения классов, а также полупрямые, содержащие только положительные и отрицательные экземпляры.
3. Отбрасываем корректно классифицированные экземпляры на данном направлении и повторяем шаги 1-3 до достижения пустого множества экземпляров или фиксированного числа итераций.

Классификация экземпляра производится по следующему алгоритму:

цикл i = 1... 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

3.4 Сопоставление весов признакам для метода опорных векторов

В работе [9] исследовались различные способы сопоставления весов признакам для метода опорных векторов. В частности, при оценке веса признака в форме

$$w_i = \ln(TF) \cdot IDF$$

были получены лучшие оценки качества классификации среди остальных подходов. В данной формуле *TF* – количество вхождений признака в документ, *IDF* – инверсная частота признака в коллекции.

В данной работе был предложен альтернативный подход к оценке веса признака

$$w_i = \ln(TF) \cdot IDF_{new}, \quad (3.4.1)$$

где IDF_{new} определяется согласно формуле

$$IDF_{new} = \sqrt{\max_{C'\in C} TF(w,C')} * IDF',$$

$$IDF' = \sqrt{\frac{|D|}{\sum_{C'\in C}\frac{TF(w,C')}{\sum_{w'\in F} TF(w',C')}}}$$

4 Результаты экспериментов

4.1 Экспериментальное сравнение алгоритмов классификации

Оценка качества алгоритмов классификации проводились на общедоступных и широко используемых тесовых коллекциях Reuters-21578, Newsgroup-20, OHSUMED и РОМИП-Legal.

При проведении экспериментов были поставлены следующие цели:

- Получить оценку качества классификации предложенных алгоритмов при решении задачи классификации с большим количеством классов в обучающей выборке и сравнить эти алгоритмы по качеству классификации с широко используемыми алгоритмами
- Получить сравнительную оценку качества классификации предложенных алгоритмов при решении задачи бинарной классификации

На представленных гистограммах использовались следующие условные обозначения алгоритмов классификации:

Bayes – метод Байеса

ModBayes – модифицированный метод Байеса (описанный в разделе 3.2)

SVM – метод опорных векторов с использованием линейного ядра

modSVM – метод опорных векторов с использованием предложенного способа сопоставления весов признакам (описанного в разделе 3.4)

modFisher – предложенный алгоритм построения нескольких разделяющих гиперплоскостей (описанный в разделе 3.3)

Общие результаты приведены на следующих двух гистограммах:

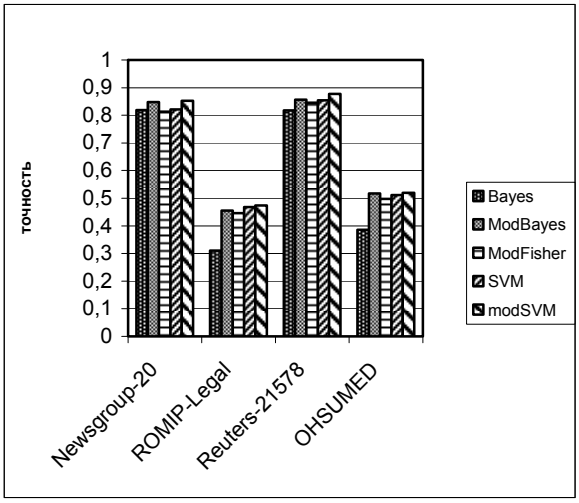


Рис.1. Общее сравнение качества алгоритмов классификации

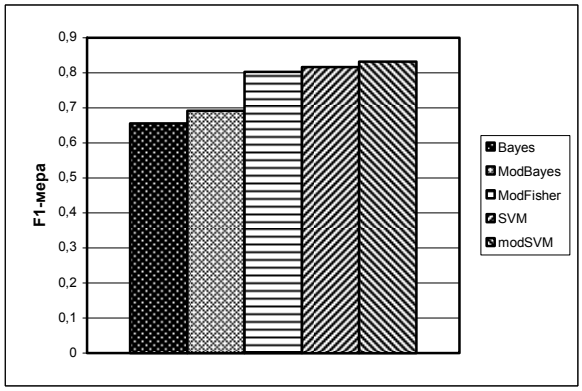


Рис.2. Сравнение качества решения задачи бинарной классификации на коллекции Newsgroup-20

По результатам экспериментов можно сделать следующие выводы:

При решении задачи бинарной классификации качество результатов предложенного алгоритма построения нескольких разделяющих гиперплоскостей сопоставимо с качеством алгоритма SVM.

Модифицированный метод Байеса при решении задачи бинарной классификации существенно проигрывает по качеству и алгоритму SVM и алгоритму ModFisher (что ожидаемо, так как парадигма класса-дополнения не работает при решении задачи бинарной классификации).

При решении задачи классификации с большим количеством классов модифицированный метод Байеса превосходит алгоритм ModFisher по качеству классификации, а также и алгоритм SVM без применения модификатора весов (3.4.1) на большинстве тестовых коллекций.

На всех тестовых наборах метод опорных векторов с примененной оценкой веса (3.4.1) превосходит по качеству классификации все рассмотренные алгоритмы.

Принимая за критерий оптимальности алгоритма соотношение качества классификации и временных затрат на обучение (эти алгоритмы обладают сравнимой вычислительной сложностью классификации), можно сделать следующие рекомендации:

В случае малого объема обучающей выборки предпочтительно использование метода опорных векторов с предложенной в разделе 3.4 схемой оценки весов признаков. Для решения задачи классификации в случае наличия в обучающей выборке большого количества классов рекомендуется применять модифицированный метод Байеса . Алгоритм ModFisher предпочтителен для решения задачи бинарной классификации при больших объемах обучающих выборок.

4.2 Результаты апробации предложенного метода

Апробация предложенного метода подхода проводилась на реальных данных Web. Производилась оценка показателей точности поиска и меры F1 на первых пятидесяти результатах поиска. Обучающая выборка составлялась с привлечением пользователей, которыми было предоставлено 30 примеров релевантных документов и 15 примеров нерелевантных документов. Множество примеров нерелевантных документов было расширено документами из нерелевантных рассматриваемым темам классов в коллекции Newsgroup-20.

Пользователями также для каждой из тем было представлено три запроса по ключевым словам, по их мнению, описывающих интересующую их тему, из которых был выбран лучший с точки зрения качества поиска. Также был сформирован второй вариант запроса по ключевым словам путем изменения запроса по ключевым словам, автоматически сгенерированного на основе обучающей выборки.

Результаты апробации приведены на следующих двух гистограммах:

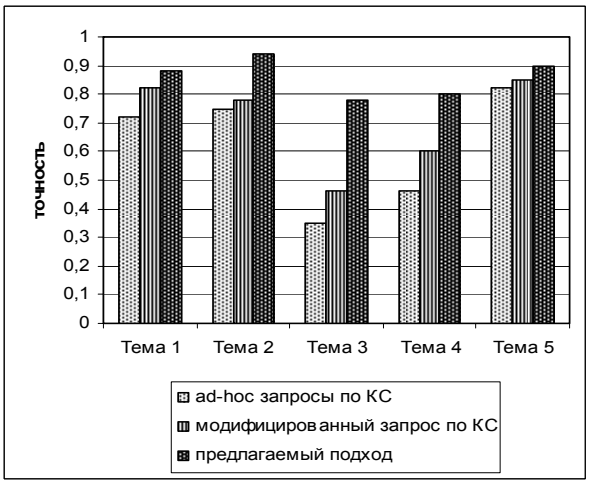


Рис.3. Сравнение точности поиска

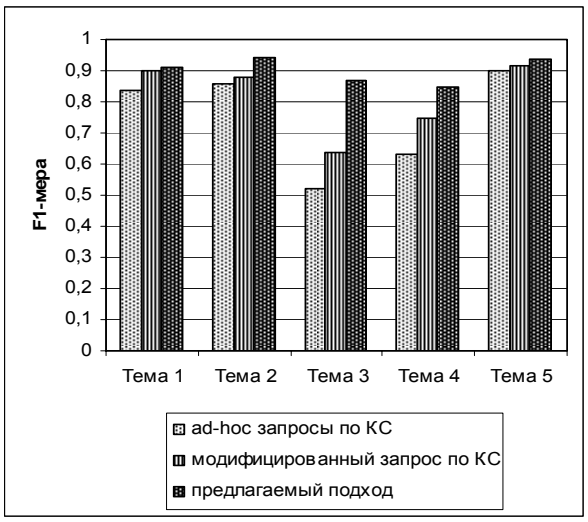


Рис.4. Сравнение качества поиска по мере F1

Полученные результаты позволяют говорить о том, что наблюдается значительное улучшение точности поиска (примерно на 16%) и существенное улучшение качества поиска, выраженное мерой F1 (примерно на 7,5%) по сравнению с обычным поиском по ключевым словам.

Благодаря использованию этапа отбора документов из Web с помощью запроса по ключевым словам в среднем удалось сократить множество анализируемых классификатором документов в среднем более чем в 10000 раз, что позволяет обеспечить приемлемое время поиска.

5 Заключение

В данной статье описывается метод периодического тематического поиска, представляющий собой композицию поиска по ключевым словам и тематической фильтрации с использованием классификаторов текстов. Исходя из требований, предъявляемых к алгоритмам классификации в рамках рассматриваемой задачи, были предложены алгоритмы классификации для решения задач бинарной классификации и задачи классификации с большим количеством классов в обучающей выборке.

Для оценки качества классификации предложенных алгоритмов мы проводили сравнение с методом опорных векторов (SVM) и методом Байеса на ряде тестовых коллекций. Результаты экспериментов говорят о сопоставимости качества классификации предложенных алгоритмов и SVM, при этом предложенные алгоритмы обладают меньшей вычислительной сложностью обучения.

Также в статье приведены результаты апробации предложенного метода периодического тематического поиска на реальных данных. Полученные результаты позволяют говорить о возможности достижения существенного улучшения качества поиска, по сравнению с методом поиска по ключевым словам.

Литература

[1] A. Barfourosh, H. Nezhad, M. Anderson, D. Perlis. Information Retrieval on the World Wide Web and Active Logic: A Survey and Problem Definition. *Techical report CS-TR-4291, UM Computer Science Department*, 2002

[2] S. Chakrabarti. Mining The Web Discovering Knowledge From Hypertext Data. *Morgan Kaufmann Publishers*, 2004.

[3] H. Chen, S. Dumais. Bringing Order to the Web: Automatically Categorizing Search. *Proceedings of ACM CHI 2000 Conference on Human Factors in Computing Systems*,volume 1, pages 145-152, 2000.

[4] O. Driori, N. Aron. Using Documents Classification for Displaying Search Results List. *Technical Report No. 2002-34 of the Leibniz Center for Research in Computer Science, Hebrew University of Jerusalem*, Jerusalem, 2002.

[5] R. Fisher. The use of multiple measurements in taxonomic problems. *Eugen.*, 7:179-188, 1936.

[6] T. Joachims. Making Large-Scale SVM Learning Practical. *Advances in Kernel Methods. Support Vector Learning*. MIT-Press, 1999.

[7] F. Johannes. A study using n-gram features for text categorization. *Technical Report OEFAI-TR-9830, Austrian Institute for Artificial Intelligence*, 1998.

[8] M. Kobayashi, K. Takeda. Information retrieval on the Web. *IBM Research Report, RT0347*, 2000.

[9] C. Liao, S. Alpha, P. Dixon. Feature Preparation in Text Categorization. *Proceedings of Australian Data Mining Conference*, Canberra, 2003

[10] J. Provost. Naive-Bayes vs. Rule-Learning in Classification of Email. *Technical Report AI-TR-99-284, The University of Texas at Austin, Department of Computer Sciences*, 1999.

[11] R. Quinlan. C4.5: Programs for Machine Learning. *Morgan Kaufmann Publishers*, 1993

[12] C. van Rijsbergen. Information Retrieval. *Butterworth's and Co.*, 1979.

[13] F. Sebastiani. Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, volume 1, pages 1-47, 2002.

[14] Y. Yang, X. Liu. A re-examination of text categorization methods. *Proceedings of International ACM Conference on Research and Development in Information Retrieval (SIGIR-99)*, pages 42-49, 1999.

On the method of periodical thematic Web search

A.V. Maksakov

This paper describes the method of periodical thematic search, based on composition of keyword search and filtering using classification algorithms. Several classification algorithms were analyzed from the point of their effectiveness in proposed method.