Оценка тематического подобия текстовых документов

В.Ю. Добрынин , В.В. Клюев, И.С. Некрестьянов
vdobr@oasis.apmath.spbu.ru, kluev@oasis.apmath.spbu.ru, igor@meta.math.spbu.ru
Санкт-Петербургский Государственный Университет

Аннотация

В статье рассматривается задача оценки тематического подобия текстовых документов в рамках заданной коллекции.

Предлагаемый алгоритм использует оригинальный подход для определения относительно небольшого приблизительного тематического окружения заданного документа. По результатам анализа полученного тематического окружения формируется множество ключевых слов, характеризующих тематику исходного документа относительно остальных документов коллекции. Построенный набор ключевых слов далее используется для вычисления относительных оценок тематической близости документов.

Экспериментальная проверка эффективности предлагаемого подхода проводилась на основе стандартного набора тестовых данных Reuters-21578.

1 ВВЕДЕНИЕ

Методы, учитывающие тематическую ориентацию документов, в последние годы применялись для решения широкого круга задач информационного поиска, таких как построение тематических индексов [4, 7], определение тематики запроса [6, 13], автоматическая классификация документов [1, 12, 8, 3], и многие другие.

В этой работе рассматривается задача поиска по документу-образцу в рамках заданной коллекции документов. Одним из возможных подходов к решению этой задачи является использование документа-образца в качестве длинного запроса и применение традиционных методов обработки запросов. К сожалению, такой подход плохо работает с документами среднего и большого размеров. Другой вариант — использовать некоторые части документа-образца для построения нескольких запросов, аппроксимирующих запрос по документу-образцу, и объединить полученные по этим запросам результаты.

Предлагаемый метод решения этой задачи основан на вычислении оценок тематического подобия двух документов. Вообще, понятие тематической близости документов относительно и определяется контекстом, в рамках которого близость оценивается. Так, например, два документа, характеризующие изменение котировок акций и изменение курса валюты соответственно, будут, вероятно, признаны тематически похожими среди случайного множества документов, но в то же время они существенно различаются в рамках узкоспециализированной экономической коллекции. Поэтому, в описываемом методе оценка тематической близости определяется не только самими документами, но и зависит от всей коллекции документов.

Известно, что словарный запас и частоты использования слов зависят от тематики [11, 6]. Мы используем это наблюдение при вычислении оценок тематической близости — мы учитываем только те слова, которые более специфичны для тематики рассматриваемого документа. Такие слова выделяются по результатам анализа аппроксимированного тематического окружения данного документа. Похожий подход использовался в работе [11] в применении к классификации потока документов.

Под влиянием ряда работ Джеральда Сэлтона [10, 9] мы полагаем, что типичный документ среднего размера затрагивает не одну тематику. Поэтому мы представляем каждый документ как последовательность частей ("параграфов"), каждая из которых отражает некоторый тематический аспект документа.

Предлагаемый метод состоит из следующих основных этапов:

- Для каждого документа определяется некоторое (относительно небольшое) множество документов, представляющее его (аппроксимированное) тематическое окружение.

- Построенные тематические окружения анализируются с целью формирования множеств ключевых слов, характеризующих тематику исходного документа относительно остальных документов коллекции.

- Полученные наборы ключевых слов используются

©Вторая Всероссийская научная конференция
ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ:
ПЕРСПЕКТИВНЫЕ МЕТОДЫ И ТЕХНОЛОГИИ,
ЭЛЕКТРОННЫЕ КОЛЛЕКЦИИ
26-28 сентября 2000г., Противно

для дальнейшего вычисления относительных оценок тематического подобия.

Экспериментальная проверка предлагаемого метода проводилась на двух англоязычных наборах тестовых данных — архиве статей из электронного журнала по комбинаторике[1] и на подмножестве документов из стандартной[2] коллекции Reuters-21578.

2 ОПИСАНИЕ МЕТОДА

Предлагаемый метод состоит из следующих основных этапов:

1. Предварительная обработка документов
2. Построение тематического окружения
3. Формирование множеств ключевых слов
4. Вычисление относительных оценок тематического подобия

Каждый из этих этапов подробно описан в последующих разделах.

2.1 Предварительная обработка документов

Предварительная обработка включает следующие операции:

- **Лексический анализ**
  - удаление разметки, т. е. элементов форматирования TEX, HTML или XML
  - стандартные операции: удаление пунктуации, цифр, преобразование всех букв в прописные, т. п.
  - удаление математических формул из документов в формате TEX

- **Исключение стоп-слов**

  Стоп-слова — это широко употребительные слова, не несущие смысловой нагрузки (например, местоимения). Мы воспользовались стандартным списком стоп-слов от известной исследовательской поисковой системы SMART[3].

- **Выделение основ слов**

  Поскольку мы практически работали с англоязычными документами, то мы воспользовались стандартным для англоязычного поиска алгоритмом Портера [2] для выделения основ слов.

- **Разбиение документа на "параграфы"**

  Целью этого разбиения является представление документа d в виде последовательности $P(d)$ тематически однородных фрагментов документа P_q.

  Отметим, что таким образом мы учитываем некоторую информацию об относительной близости термов в тексте документа, т. е. используем некоторую разновидность методов анализа локального контекста [5].

  Выделение из документа частей, затрагивающих различные тематики, — относительно трудоемкая задача [9]. В рамках этой работы мы используем эвристический подход, основанный на желаемом размере получаемого фрагмента L.

  Мы выбрали такую эвристику, исходя из следующих соображений. Использование сильно отличающихся в размере параграфов имеет ряд недостатков: словарь коротких параграфов значительно беднее словаря длинных параграфов, благодаря чему термы в профайлах коротких параграфов в среднем имеют больший вес. Все это приводит к излишним помехам при вычислении оценки близости.

  Параграфы строятся согласно следующим правилам:

  Авторские параграфы. Мы предполагаем, что автор обычно разбивает текст на параграфы, учитывая их тематику. Тем самым, такое разбиение содержит неявные экспертные оценки. Опираясь на это наблюдение, мы используем авторские параграфы в качестве основы нашего разбиения.

  Желаемый размер. Для получения параграфов желаемого размера мы либо объединяем несколько подряд идущих коротких авторских параграфов, либо разбиваем слишком большие параграфы на части так, чтобы размер получающихся параграфов не отличался от желаемого размера L более чем на 20 процентов.

- **Построение профайлов параграфов**

  Мы используем векторную модель, в рамках которой профайлы представляют собой векторы. В данном случае для каждого "параграфа" P_q формируется так называемый tf-профайл, сопоставляющий каждому терму t частоту его встречаемости в данном "параграфе" $\mathrm{tf}_q(t)$.

2.2 Построение тематического окружения документа

Тематическое окружение документа используется для выявления тех особенностей, которые характеризуют тематическую ориентацию рассматриваемого документа d_0 относительно рассматриваемого набора документов C. Поэтому важно, чтобы доля документов, тематически близких данному документу была в построенном тематическом окружении выше, чем в рамках всей коллекции C.

[1] Electronic Journal of Combinatorics — http://www.combinatorics.org

[2] Коллекция Reuters-21758 доступна по адресу http://www.research.att.com/~lewis/reuters21578.html.

[3] ftp://ftp.cs.cornell.edu/pub/smart/

В тематическое окружение $T(d_0)$ документа d_0 включаются все документы d_c, которые признаются тематически подобными заданному документу d следующим алгоритмом:

1. Формируются два виртуальных документа, являющиеся конкатенациями документов d_0, d_c и d_c, d_0 соответственно.

2. Для каждого из виртуальных документов оценивается мера близости для каждой пары соседних параграфов $(k, k+1)$:

   - Для каждого параграфа P_q формируется стандартный tf·idf–профайл, т. е. вектор, в котором каждому терму t сопоставляется вес

     $$W_q(t) = \frac{\mathrm{tf}_q(t)\log(\frac{N}{n(t)})}{\sqrt{\sum\limits_{k\in D}(\mathrm{tf}_q(k)\log(\frac{N}{n(k)}))^2}}$$

     где $n(t)$ обозначает число параграфов виртуального документа, содержащих терм t, N — общее число параграфов, а D — общий словарь виртуального документа.

   - Используя стандартную для векторной модели меру косинуса, для каждого параграфа P_i определяется наиболее близкий к нему параграф $P_j,\ j > i$:

     $$w(i,j) \stackrel{\mathrm{def}}{=} (W_i, W_j) = \min_k\,(W_i, W_k)$$

   - Мера близости соседних параграфов k и $k+1$ определяется как

     $$\mathrm{sim}(k, k+1) \stackrel{\mathrm{def}}{=} \sum_{i\le k}\sum_{j\ge k+1} w(i,j).$$

3. Последовательность значений $\mathrm{sim}(k, k+1)$ используется для разбиения виртуального документа на тематически однородные группы параграфов — границы групп соответствуют тем значениям k, при которых значение функции sim значительно меньше значений этой функции в соседних точках, т. е.

   $$\mathrm{sim}(k-1) > \alpha\cdot\mathrm{sim}(k) < \mathrm{sim}(k+1)$$

4. Окончательно, документ d_c считается тематически близким документу d_0, если для обоих виртуальных документов ни одна из границ между тематически однородными группами параграфов не совпала с границей между документами d_0 и d_c.

2.3 Формирование множеств ключевых слов

Множество ключевых слов $K(d_0)$ для документа d_0 — это подмножество термов, встречающихся в d_0, которое характеризует тематику этого документа. Формирование множества происходит на основе сравнительного анализа статистики использования термов в рамках коллекции в целом и в рамках тематического окружения заданного документа.

Исходя из предположения о том, что доля релевантных d_0 документов выше в его тематическом окружении $T(d_0)$, чем в коллекции в целом, мы полагаем, что характерные для тематики документа d_0 термы также встречаются в построенном тематическом окружении чаще, чем в среднем.

Для формирования множества ключевых слов документа d_0 используется следующий подход. Для каждого терма t документа d_0 вычисляется вероятность появления этого терма в случайно выбранном документе из тематического окружения $T(d_0)$. В множество ключевых слов включаются те термы из документа d_0, для которых эта вероятность значительно превосходит вероятность появления данного терма в документе, случайно выбранном из всей коллекции C, т. е. таких $t \in d_0$, что

$$\frac{|\{d : d \in T(d_0), t \in d\}|}{|T(d_0)|} > \beta\cdot\frac{|\{d : d \in C, t \in d\}|}{|C|}$$

2.4 Вычисление относительных оценок тематического подобия

Построенное множество ключевых слов $K(d_0)$, характеризующих тематику документа d_0 относительно рассматриваемой коллекции, используется для вычисления относительной оценки степени тематической близости $tsim(d_0, d)$ между d_0 и другими документами коллекции.

Введем следующие обозначения. Пусть $K_q(d)$ обозначает множество ключевых слов документа d, которые встречаются в его параграфе P_q. Пусть множество

$$E(d) \stackrel{\mathrm{def}}{=} \{\{x, y\} : x, y \in K_q(d), q \in P(d)\}$$

обозначает множество всех пар ключевых слов документа d, которые хотя бы раз совместно встречаются в одном параграфе документа d, а

$$E(d|d') \stackrel{\mathrm{def}}{=} \{\{x, y\} : \{x, y\} \in E(d), x, y \in K(d')\}$$

есть сужение E(d) на множество пар ключевых слов документа d'.

Тогда оценка тематической близости определяется как

$$tsim(d_0, d) \stackrel{\mathrm{def}}{=} \frac{|E(d|d_0) \cap E(d_0)|}{|E(d|d_0) \cup E(d_0)|}$$

3 ЭКСПЕРИМЕНТАЛЬНЫЕ РЕЗУЛЬТАТЫ

Экспериментальная проверка предлагаемого метода проводилась на двух англоязычных наборах тестовых данных:

- архиве статей из электронного журнала по комбинаторике[4]

- на подмножестве документов из стандартной[5] коллекции Reuters-21578.

[4]Electronic Journal of Combinatorics — `http://www.combinatorics.org`

[5]Коллекция Reuters-21578 доступна по адресу `http://www.research.att.com/Ĩewis/reuters21578.html`.

3.1 Эксперименты с коллекцией статей по комбинаторике

Эта коллекция состоит из 250 научных статей по различным разделам комбинаторики в формате TEX. К сожалению, эта коллекция специально не предназначена для экспериментов в области информационного поиска и для нее не существует экспертных оценок тематической близости для каждой пары статей.

Мы привлекли к оценке результатов эксперта в области комбинаторики, но он, естественно, не мог выставить оценки тематической близости для всех возможных пар документов. Поэтому мы ограничились рядом неформальных экспериментов с этим набором данных.

В частности, мы сопоставляли результаты поиска по ключевым словам, используя систему Excite for Web Servers, и по документу-образцу, используя предложенный в этой статье метод.

В рассматриваемой коллекции присутствуют две статьи разных авторов с одинаковым названием "The Problem of the Kings", посвященные вычислению числа способов размещения максимального числа неатакующих королей на шахматной доске размера $2m \times 2n$.

По запросу, совпадающему с названием этих статей, первые шесть найденных Excite статей получили оценки вероятности релевантности в 74, 72, 54, 52, 28, 26 процентов соответственно. Однако только первая, вторая и четвертая из них были признаны истинно релевантными экспертом.

При использовании в качестве документа-образца первой статьи с названием "The Problem of the King" предлагаемый метод сформировал список документов, в котором все 3 релевантных документа, найденные Excite, были расположены на первых трех позициях. При использовании же в качестве документа-образца второй статьи были обнаружены не только все три упомянутых ранее релевантных документа, но также еще три новых релевантных документа (комбинаторика, связанная с шахматами), которые были помещены на 4, 6 и 7 места.

Этот пример, в частности, иллюстрирует тот факт, что построенное нами отношение тематической релевантности не транзитивно, т. е. если документы A и B тематически релевантны документу C, то это не означает, что они тематически релевантны друг другу.

3.2 Эксперименты с Reuters

Коллекция Reuters–21578 является одним из широко используемых стандартных наборов тестовых данных в области информационного поиска. К сожалению, эта коллекция также не содержит всей необходимой информации для того, чтобы оценить качество поиска по документу-образцу. Тем не менее, доступная информация может быть использована для проведения более формальной оценки эффективности предлагаемого подхода.

3.2.1 Характеристика набора данных

Коллекция Reuters-21578 содержит документы преимущественно экономической тематики, и примерно половине документов эксперты приписали одну или несколько тем (всего имеется 135 различных тем). Мы выбрали те документы, которым приписана хотя бы одна тема, и длина которых не менее 3000 байт. Таким образом, мы отобрали 411 документов на 72 темы.

Критерий	min	max	среднее
размер документа (байт)	3001	13470	3968
размер темы (документов)	1	84	10.3
число документов релевантных документу d, $R(d)$	1	203	55

Таблица 1: Характеристика тестового набора данных.

Поскольку этот набор данных не содержит явных оценок тематической релевантности между документами, то мы использовали эвристический подход. При оценке точности и полноты ответа на запрос мы считали найденный документ релевантным документу-образцу, если хотя бы одна тема документа-образца присутствует среди тем найденного документа. Конечно, такой критерий довольно приблизителен, но все же он достаточен для относительной оценки эффективности метода.

Более подробная информация о тестовом наборе изложена в таблице 1.

3.2.2 Качество тематических окружений

Используя тематические окружения, мы предполагаем, что процент релевантных документов в них выше, чем в среднем в коллекции. Выполнение этого предположения и является основным критерием ($Satisfaction$) качества построенных окружений. Важной количественной характеристикой является также *Goodness* — усредненное соотношение процентов релевантных документов в рамках тематического окружения и в рамках всей коллекции.

На практике качество получаемых тематических окружений зависит от параметров экспериментов (а именно от L и α). Тем не менее, в большинстве экспериментов сделанное предположение обычно соблюдалось в более чем 97% случаев и коэффициент превышения (*Goodness*) изменялся от 5 до 8 (максимально возможное значение *Goodness* равно 15.2).

Так, в эксперименте с параметрами ($L = 100$, $\alpha = 2$) размер тематических окружений в среднем составил 37 документов, изменяясь от 0 (в трех случаях) до 127, средняя полнота окружений составила 35%, а точность — 55%. Процент релевантных документов оказался выше, чем в среднем, в 403 случаях из 411 (т. е. $Satisfaction > 97\%$) и средний коэффициент превышения составил 5.55. Из 8 неудач в трех случаях соотношение было очень близко к единице и еще в трех были построены пустые тематические окружения.

3.2.3 Эффективность метода ранжирования

Эта группа экспериментов имела своей целью оценить эффективность предложенного метода вычисления относительных оценок релевантности.

Очевидно, что качество вычисленных оценок релевантности сильно зависит от входных данных, т. е. от качества использовавшихся тематических окружений. Поэтому, при оценке эффективности предложенного метода

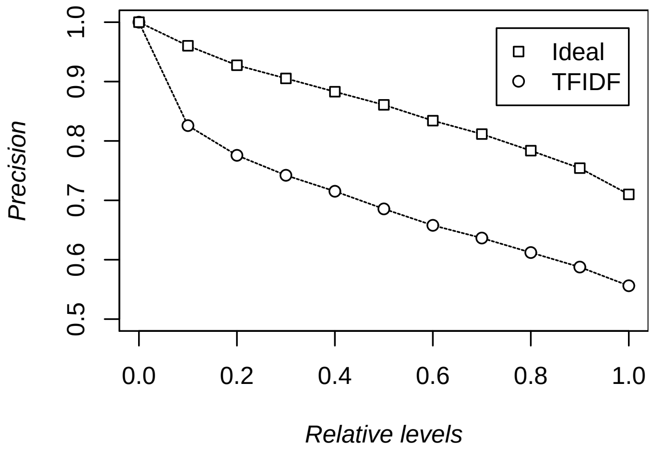

Рис. 1: Эффективность метода ранжирования.

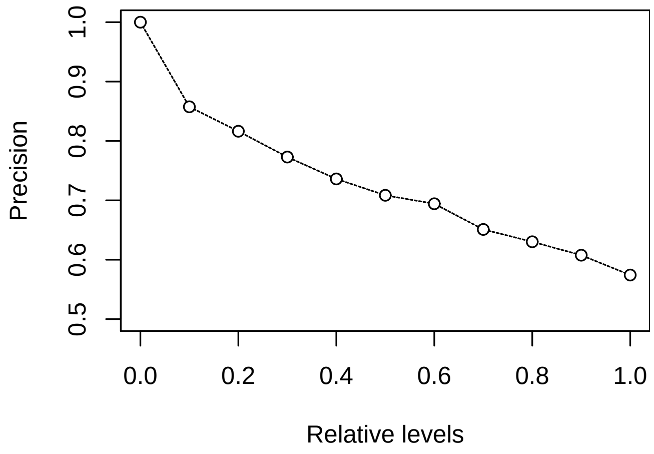

Рис. 2: Общая эффективность.

ранжирования мы использовали набор “идеальных” тематических окружений, каждое из которых содержит только релевантные центроиду документы.

Для сравнения мы используем результаты ранжирования стандартным методом, использующим tf·idf подход для определения весов термов в профайлах документов и скалярное произведение профайлов для вычисления ранга [2].

Поскольку мы заинтересованы в общей оценке поведения метода, то нас интересуют усредненные показатели. В качестве одного из рассматриваемых нами показателей выступает усредненное значение $P@N$ — точности на уровне N, т. е. точности при рассмотрении N первых документов в получаемых ранжированных списках. Однако, поскольку общее число релевантных документов сильно варьируется от случая к случаю, более естественным кажется проводить усреднение не по абсолютным, а по относительным уровням. Так, мы измеряем усредненное значение $R - precision$ — значение точности на уровне $R(d)$.

Аналогичным образом, мы усредняем график зависимости точности от числа возвращаемых результатов по относительным уровням (на графиках используется обозначение “relative levels”), т.е. пропорционально $R(d)$.

Результаты сравнения эффективности предлагаемого метода ранжирования и стандартного подхода изображены в виде такого графика на рисунке 1.

3.2.4 Общая эффективность метода

Для оценки общей эффективности метода мы использовали те же критерии, что и для оценки эффективности процедуры ранжирования.

В эксперименте с параметрами ($L = 100$, $\alpha = 2$, $\beta = 1$) среднее значение $R - precision$ составило 57%. График усредненной (по относительным уровням) зависимости точности от числа возвращенных документов изображен на рисунке 2. Напомним, что, согласно используемому нами критерию истинной релевантности, каждому документу считалось релевантным около 13% всех документов.

По сравнению с результатами ранжирования с использованием tf·idf подхода в этом эксперименте также получились лучшие результаты. Хотя, конечно, они заметно хуже результатов ранжирования на основе идеальных тематических окружений.

Размер параграфа L	20	50	100	200
средний размер	24.8	35.7	36.7	37.3
наличие превышения (%)	96.8	97.1	98.1	96.4
коэффициент превышения	4.28	5.09	6.05	5.67

Таблица 2: Зависимость качества тематических окружений от длины параграфа.

3.2.5 Влияние эвристических параметров

Описанный метод использует несколько параметров, изменение которых влечет изменение получаемых результатов. Мы провели ряд предварительных экспериментов с целью определить эвристики для выбора некоторых параметров.

- **Оптимальная длина параграфа L**

  Для того, чтобы оценить влияние длины параграфа на качество получаемых результатов (а именно на качество получаемых тематических окружений) мы провели ряд экспериментов, зафиксировав все параметры ($\alpha = 2, \beta = 1$), кроме длины параграфа L. Результаты просуммированы в таблице 2.

  Они демонстрируют, что оптимальная длина параграфа в данном случае составляет около 100 слов. Эти цифры нельзя рассматривать как абсолютные, поскольку идеальный размер параграфа скорее всего зависит от рассматриваемой коллекции (и, возможно, от значений других параметров). Тем не менее, хочется отметить относительную стабильность результатов в относительно широком диапазоне длин параграфов — от 50 до 200. Это дает надежду, что для получения приличных результатов может быть достаточно относительно грубой эвристики для выбора размера параграфа.

- **Влияние параметра β**

β	R-precision	P@10	Размер списка ключевых слов
1.0	0.71	0.86	66.8 (7 – 200)
1.1	0.69	0.855	59.4 (0 – 196)
1.2	0.67	0.853	53.9 (0 – 195)
1.5	0.58	0.792	41.6 (0 – 180)
1.7	0.50	0.725	35.9 (0 – 177)
2.0	0.43	0.629	30.2 (0 – 168)

Таблица 3: Влияние параметра β на эффективность метода ранжирования.

Параметр β определяет какие ключевые слова будут отобраны в качестве характеризующих коллекцию. Увеличивая значение параметра, мы уменьшаем размер построенных описаний и, тем самым, сокращаем вычислительные затраты при вычислении оценок тематической близости. С другой стороны, это влечет получение бедных (иногда пустых) описаний коллекций и, соответственно, снижение качества вычисляемых оценок.

Мы провели ряд экспериментов ($L = 100, \alpha = 2$) с разными значениями параметра β. Измеряемые характеристики — R-precision, точность среди первых 10 документов, а также средний, минимальный и максимальный размер полученных описаний. Результаты в таблице 3. Как показывает практика, увеличивая значение β, зачастую удается сократить размер описаний, незначительно ухудшая качество результатов.

4 ОБСУЖДЕНИЕ

Несмотря на относительно неплохие результаты экспериментов, описанный подход имеет ряд слабых мест:

Низкое качество тематических окружений: Как показали эксперименты с использованием идеальных тематических окружений, общая эффективность метода может быть значительно выше.

Высокая трудоемкость: Трудоемкость описанного подхода определяется огромной трудоемкостью процедуры построения тематических окружений. Из-за этой особенности метод прямо не применим к коллекциям большого размера.

Отсутствие поддержки коротких документов: На данный момент метод работает только с относительно длинными (больше нескольких параграфов) документами.

Зависимость от распределения документов: Если документы распределены по тематикам сильно неравномерно, то для относительно мало представленных тематик результаты значительно хуже, чем для тематик, представленных значительным числом документов.

Для устранения этих недостатков, а также определения способов выбора идеальных значений для параметров метода, требуются дополнительные исследования.

Например, для того, чтобы работать с содержащими короткие документы коллекциями, кажется разумным использовать какой-нибудь гибридный подход, обрабатывающий короткие документы специальным образом.

Для применения метода к коллекциям большого размера можно попробовать выделять из большой коллекции относительно небольшое множество документов так, чтобы оно было тематически представительно для всей коллекции в целом. Далее, можно использовать это множество для построения описаний тематик, а тематическую принадлежность не попавших в это множество документов можно оценивать, анализируя распределение в них ключевых слов документов из множества.

Мы также планируем исследовать применимость построенных подобным образом описаний тематических окружений к решению других задач, включая автоматическую классификацию потоков документов и кластеризацию документов по тематической похожести.

5 ЗАКЛЮЧЕНИЕ

В работе предложен новый метод вычисления тематического подобия и применения этих оценок к задаче поиска по документу-образцу.

Предлагаемый алгоритм использует оригинальный подход для определения относительно небольшого приблизительного тематического окружения заданного документа. По результатам анализа полученного тематического окружения формирует множество ключевых слов, характеризующих тематику исходного документа относительно остальных документов коллекции. Полученный набор ключевых слов и используется для вычисления оценок тематической близости документов.

Результаты экспериментов демонстрируют перспективность предлагаемого подхода. Хотя предложенный алгоритм имеет ряд слабых мест, мы надеемся, что дополнительные исследования помогут от них избавиться и улучшить общую эффективность.

Список литературы

[1] И.Е. Кураленок and И.С. Некрестьянов. Автоматическая классификация документов с использованием семантического анализа. In *Труды первой всероссийской научно-методической конференции “Электронные библиотеки: перспективные методы и технологии, электронные коллекции”*, pages 86–96, Санкт-Петербург, October 1999.

[2] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval.* ACM Press, 1999.

[3] Douglas L. Baker and Andrew Kachites McCallum. Distributional clustering of words for text classification. In *Proceedings of the SIGIR'98*, pages 96–103, 1998.

[4] Soumen Chakrabarti, Martin van den Berg, and Byron Dom. Focused crawling: A new approach to topic-specific web resource discovery. In *Proc. of the WWW-8*, May 1999.

[5] Bruce W. Croft and Jinxi Xu. Query expansion using local and global document analysis. In *Proc. of the SIGIR'96*, pages 4–11, 1996.

[6] Igor Kuralenok, Vladimir Dobrynin, Igor Nekrestyanov, Mikhail Bessonov, and Ahmed Patel. Distributed search in topic-oriented document collections. In *Proc. of World Multiconference on Systemics, Cybernetics and Informatics (SCI'99)*, volume 4, pages 377–383, August 1999.

[7] Igor Nekrestyanov, Tadhg O'Meara, and Ekaterina Romanova. Building topic-specific collections with intelligent agents. In *Proc. of Sixth International Conference on Intelegence in Services and Networks (IS&N'99)*, Barcelona, Spain, April 1999.

[8] Ron Papka and James Allan. Document classification using multiword features. In *Proc. of the CIKM'98*, pages 124–131, November 1998.

[9] Gerald Salton, James Allan, and Amit Singhal. Automatic text decomposition and structuring. *Information Processing & Management*, 32(2):127–138, 1996.

[10] Gerald Salton, Amit Singhal, Mandar Mitra, and Chris Buckley. Automatic text decomposition and summarization. *Information Processing & Management*, 33(2):193–208, 1997.

[11] Amit Singhal, Mandar Mitra, and Chris Buckley. Learning routing queries in a query zone. In *Proc. of the SIGIR'97*, pages 25–32, July 1997.

[12] Raymie Stata, Krishna Bharat, and Farzin Maghoul. The term vector database: fast access to indexing terms for web pages. In *Proc. of the WWW-9*, May 2000.

[13] Atsushi Sugiura and Oren Etzioni. Query routing for web search engines: Architecture and experiments. In *Proc. of the WWW-9*, May 2000.