МЕТОД ПРЕДВАРИТЕЛЬНОЙ ОЦЕНКИ ЭФФЕКТИВНОСТИ СЕМАНТИЧЕСКИХ МЕТОДОВ ОБРАБОТКИ ТЕКСТОВОЙ ИНФОРМАЦИИ.

Кураленок Игорь Евгеньевич
Санкт-Петербургский Государственный Университет
19906 Санкт-Петербург, Университетская д.7/9

Аннотация

В нашей работе обсуждается возможность построения предварительной оценки эффективности семантических методов без применения тестовых коллекций и рассматривается пример такого построения. Эксперименты, проведенные на достаточно больших объемах тестовых данных, подтверждают правомочность предложенного подхода, что позволяет рассчитывать на дальнейшее развитие этой области.

1. Введение

Актуальность области эффективной обработки текстовой информации трудно переоценить. Этот факт связан, в первую очередь, с взрывным ростом объемов информации такого рода и непрерывным совершенствованием технологий доступа к ним. Развитие эффективных методов обработки в значительной мере осложняется сложностью процедуры их сравнения.

На сегодняшний момент для решения этой задачи используется метод *тестовых коллекций*. Этот метод заключается в сравнении результатов работы исследуемой схемы на заранее определенных данных с оценками экспертов на тех же данных. В результате сравнения получается одна-двух кретериальная оценка эффективности.

Несмотря на достаточную точность оценки реальной эффективности при условии удачного подбора тестовых данных, описанный подход не лишен существенных недостатков. Так, например, очень сложно ``удачно'' подобрать данные. Этой проблеме в последнее время уделяется большое внимание[6]. Также достаточно сложен и ресурсоемок сам процесс проверки, как на стадии построения данных, так и при проведении непосредственного тестирования -- тестовые данные могут содержать сотни тысяч документов[8,6].

Вопрос трудоемкости проверки особенно актуален при построении методов, зависящих от случайных данных [2,5,3], когда при наличии нескольких случайных выборок необходимо выбрать наиболее эффективную. К тому же тестовые данные не всегда доступны. Так, в случае русского языка этот факт сильно осложняет исследования в области проверки адаптации статистических методов, тестировавшихся на других языках.

В нашей статье мы исследуем возможность построения метрики на множестве методов обработки текстовой информации, по которой можно было бы построить оценку эффективности с меньшими временными затратами на тестирование и без привязки к ручной работе, необходимой при построении тестовых данных. Построение такой метрики на всем множестве методов требует дополнительных исследований по обобщению схем текстовой обработки, поэтому на данном этапе мы ограничились классом методов, построенном в п.2.2.

На сегодня такой подход к области сравнения методов обработки текстовой информации представляется нам сравнительно новым. Несмотря на достаточную сложность задачи, результаты наших экспериментов показывают правомочность указанного подхода, что позволяет надеяться на дальнейшее развитие этой области.

2. Постановка задачи

Область обработки текстовой информации (Information Retrieval) с каждым днем порождает новые и все более сложные задачи. Несмотря на это, можно выделить несколько классических проблем, таких как поиск по коллекции, маршрутизация запроса, фильтрация, классификация, кластеризация и т.п.

Несмотря на различия формулировок этих задач, в результате нашего анализа наиболее распространенных на сегодня методов оказалось, что в ходе решения рано или поздно встает проблема ранжирования одних элементов относительно других. При этом в качестве элементов выступали вектора свойств (feature)[1] -- образы того или иного параметра задачи (документа, запроса, и т.п.).

К тому же, в ряде исследований отмечалось, что методы, основанные на одной и той же схеме ранжирования, показывают схожую относительную эффективность при решении *различных* задач. Так, например, известно, что метод поиска, основанный на частотной модели (TF), в среднем менее эффективен, чем метод, построенный с применением LSI [4]. Та же картина наблюдается и при решении задачи фильтрации [10].

Таким образом, можно сделать вывод: в оценке конечной эффективности метода решения той или иной задачи большую роль играет оценка эффективности используемой схемы ранжирования. Для дальнейших рассуждений определим задачу ранжирования более строго:

Задача 1 (Сравнение образов) Для двух заданных образов, требуется построить оценку тематической близости $rank(a,b)$ так, что при $rank(a,b)=1$ тематики прообразов a и b полностью совпадали и для того чтобы выполнялось $rank(a,b)>rank(a,c)$ было необходимо и достаточно того, что прообраз b был ближе по тематике к прообразу a, чем прообраз c.

К сожалению, несмотря на расплывчатость этой формулировки, на сегодняшний момент поставить задачу более четко представляется затруднительным в связи с тем, что не существует объективных формулировок конечных задач. Так, например, в задаче поиска присутствует субъективное понятие релевантности, в фильтрации -- тематичности и т.п.

Как было отмечено выше, эффективность конечного метода решения той или иной задачи в значительной степени зависит от эффективности используемой схемы ранжирования, поэтому в нашей работе мы поставили себе задачу исследования именно этого фактора.

2.1 Понятие сравнительной эффективности

Очевидно, что эффективность метода решения любой задачи зависит от тестовых данных. Мы ставим своей целью оценку средней эффективности, то есть эффективности при условии случайных данных. Для этого мы строим метрику на множестве схем ранжирования так, что по любым двум известным эффективностям и расстояниям между ними и исследуемым методом, указанную оценку можно построить по формуле:

$$ef_c = \begin{cases} ef_a - \|a-c\|, \|a-c\|+\|b-c\| > \|a-b\| \wedge \|a-c\| \le \|b-c\| \\ ef_a + \|a-c\| \end{cases} \quad (1)$$

где ef_a -- эффективность метода , и известно, что эффективность метода b выше. В качестве такой метрики мы рассмотрели *сравнительную эффективность* двух методов - модуль разности средних эффективностей этих методов.

В том случае если понятие эффективности многокретериально, как в случае поиска, в качестве средней эффективности можно рассмотреть интегральное обобщение этих критериев. В частности, для задачи поиска:

[1] Наиболее простым видом свойств являются слова, содержащиеся в документе, однако в последнее время наблюдается тенденция к усложнению смысла свойств. Поэтому далее мы будем пользоваться именно этим термином.

$$e(f) = \int_0^1 p_f dr_f \quad (2)$$

Таким образом, конечная форма сравнительной эффективности может быть выписана:

$$\Delta(f, g) = \left| \int_0^1 p_f dr_f - \int_0^1 p_g dr_g \right| \quad (3)$$

2.2 Семантические методы

Как отмечалось выше, для построения оценки сравнительной эффективности во всем классе схем ранжирования необходимы дополнительные исследования в области обобщения этих схем. Поэтому в нашей работе мы ограничились рассмотрением методов ранжирования, действующих по следующему алгоритму:

1. Выделение всех свойств и построение общего словаря T

2. Преобразование $D \to R^{|T|}$ по формуле

$$\vec{d}_i = (d_{i1}, d_{i2}, \ldots, d_{i|T|})$$
$$d_{ij} = \begin{cases} f(d_i, t_k), t_k \in d_i \\ 0, t_k \notin d_i \end{cases} \quad (4)$$

3. Сравнение полученных образов и вычисление ранга

$$rank(d_1, d_2) = \vec{d}_1 X \vec{d}_2 \quad (5)$$

где X -- матрица размерности $|T| \times |T|$. Функции f могут различаться для первого и второго аргумента. Далее мы будем обозначать f_1-- функцию первого аргумента, f_2 -- второго. Эти функции выбираются таким образом, чтобы отражать встречаемость и значимость терма в той или иной структуре, поэтому мы будем называть их *функциями значимости*.

Можно заметить, что предложенный класс очень широк. В него входят большое количество популярных методов, таких как Boolean, TF, TFIDF[11], LSI[12], PLSI[5], и другие[2,1,3,13,7,9]. Так как большинство схем, входящих в это семейство, являются так называемыми *семантическими* методами, и, к тому же, все широко распространенные семантические методы подчиняются этому алгоритму, все дальнейшие рассуждения относятся, в первую очередь, именно к ним[2].

3. Метод оценки сравнительной эффективности

Несмотря на определенность алгоритма работы схем ранжирования, этой информации недостаточно для построения оценки сравнительной эффективности. Для этого необходимо построить связь между объективными параметрами, такими, как близость параметров схемы, и субъективными, такими, как конечная эффективность.

3.1 Гипотеза соответствия

В основу предлагаемого метода легло предположение о зависимости сравнительной эффективности двух методов от вероятности одинаковой оценки сравнительной близости двух документов. Далее в статье мы будем называть это предположение *гипотезой соответствия*.

Гипотеза 1 (Гипотеза соответствия) Сравнительная эффективность методов сравнения образов коррелирует с вероятностью совпадений знаков сравнительной оценки близости соответствующих прообразов по всей совместной базе.

[2] Мы придерживаемся следующего определения семантических методов: семантические методы это такие методы обработки текстовой информации, которые учитывают семантические связи между различными термами.

$$\rho(m_1, m_2) = P\{(rank_{m_1}(a,b) - rank_{m_1}(a,c)) \cdot (rank_{m_2}(a,b) - rank_{m_2}(a,c)) > 0 \quad (6)$$

Не смотря на кажущуюся очевидность этого предположения (чем чаще схемы дают одинаковые ответы на вопрос ``кто ближе'' тем более близки они по эффективности), строгое доказательство этого факта провести невозможно, в связи с тем, что понятие эффективности -- субъективно, в то время как значение ρ может быть вычисленно на любом наборе документов строго.

3.2 Построение оценки сравнительной эффективности

В этом пункте мы выведем оценку сравнительной эффективности методов сравнения образов на основе гипотезы соответствия и ограничений на структуру рассматриваемых методов, наложенную в п. 2.2.

Подставим в форму (6) вид рассматриваемых схем (5):

$$\begin{aligned} &\rho(m_1, m_2) = P\{(\vec{a}^T X_1 \vec{b} - \vec{a}^T X_1 \vec{c}) \cdot (\vec{a}^T X_2 \vec{b} - \vec{a}^T X_2 \vec{c}) > 0\} \Leftrightarrow \\ &\rho(m_1, m_2) = P\{\vec{a}^T X_1 (\vec{b} - \vec{c}) \cdot \vec{a}^T X_2 (\vec{b} - \vec{c}) > 0\} \end{aligned} \quad (7)$$

Проведя замену переменных $\vec{x} = \frac{\vec{a}}{\|\vec{a}\|}, \vec{y} = \frac{(\vec{b} - \vec{c})}{\|\vec{b} - \vec{c}\|}$ и учитывая то, что $\|\vec{b} - \vec{c}\| \cdot \|\vec{a}\|$ не влияют на знак рассматриваемого выражения, получим:

$$\rho(m_1, m_2) = P\{\vec{x}^T X_1 \vec{y} \cdot \vec{x}^T X_2 \vec{y} > 0\} \quad (8)$$

В том случае, когда функции значимости равны и симметричны относительно нуля, значения x и y можно рассмотреть как независимые случайные величины, равномерно распределенные по единичной сфере с центром в $\vec{0}$. Далее можно разделить указанное событие на два непересекающихся:

$$\rho(m_1, m_2) = P\{\vec{x}^T X_1 \vec{y} > 0 \wedge \vec{x}^T X_2 \vec{y} > 0\} + P\{\vec{x}^T X_1 \vec{y} < 0 \wedge \vec{x}^T X_2 \vec{y} < 0\} \quad (9)$$

К сожалению, данная модель, несмотря на свою кажущуюся простоту, не поддается прямому анализу без дополнительных предположений. Поэтому мы рассмотрели только тот случай, когда документы b и c отличаются ровно на один терм, причем этот терм равномерно распределен по словарю. Таким образом мы переходим к следующей системе:

$$\begin{aligned} &\rho(m_1, m_2) = \frac{1}{|T|} \sum_{i=0}^{|T|} \left(P\{\vec{x}^T \vec{x}_1^i > 0 \wedge \vec{x}^T \vec{x}_2^i > 0\} + P\{\vec{x}^T \vec{x}_1^i < 0 \wedge \vec{x}^T \vec{x}_2^i < 0\} \right) \\ &\rho(m_1, m_2) = \frac{1}{|T|} \sum_{i=0}^{|T|} \left(P\{\vec{x}_1^{iT} \vec{x} > 0 \wedge \vec{x}_2^{iT} \vec{x} > 0\} + P\{\vec{x}_1^{iT} \vec{x} < 0 \wedge \vec{x}_2^{iT} \vec{x} < 0\} \right) \end{aligned} \quad (10)$$

Рассмотрим сечение нашего пространства плоскостью $(\vec{x}_1^i, \vec{x}_2^i)$, проходящей через начало координат. Можно заметить, что эта плоскость будет ортогональна пересечению гипер-плоскостей $\vec{x}_1^{iT} \vec{x} = 0$ и $\vec{x}_2^{iT} \vec{x} > 0$. Сечение единичной сферы представляет собой окружность, из чего следует, что вероятность

$P\{\vec{x}_1^{iT} \vec{x} > 0 \wedge \vec{x}_2^{iT} \vec{x} > 0\} + P\{\vec{x}_1^{iT} \vec{x} < 0 \wedge \vec{x}_2^{iT} \vec{x} < 0\}$ для векторов из выбранного сечения составляет.

$\frac{1}{\Pi}\left(\Pi - \arccos\left(\frac{(\vec{x}_1^i, \vec{x}_2^i)}{\|\vec{x}_1^i\| \cdot \|\vec{x}_2^i\|}\right)\right)$ Для любого движения вдоль вектора $\vec{r} \in \left\{x \middle| \vec{x}_1^i \vec{x} = 0 \cap \vec{x}_2^i \vec{x} = 0\right\}$ эта ситуация будет сохранятся, из чего можно сделать вывод, о том, что полученная формула верна для любых векторов из единичной сферы. Поэтому мы можем переписать (10):

$\left(\vec{x}_1^{\,i}, \vec{x}_2^{\,i}\right)$

$$\rho(m_1, m_2) = \frac{1}{\Pi \cdot |T|} \sum_{i=0}^{|T|} \left(\Pi - \arccos\left(\frac{\left(\vec{x}_1^{\,i}, \vec{x}_2^{\,i}\right)}{\left\|\vec{x}_1^{\,i}\right\| \cdot \left\|\vec{x}_2^{\,i}\right\|} \right) \right) \tag{11}$$

Особого внимания заслуживает случай знакопостоянных функций значимости, так как именно такой вид функций наиболее распространен. В этом случае нельзя рассматривать x и y как независимые случайные вектора, равномерно распределенные по сфере, так как это неверно на любом множестве документов. Несмотря на это все проведенные выше рассуждения верны. Рассмотрим сечения пространства двумерной плоскостью $\left(\vec{x}_1^{\,i}, \vec{x}_2^{\,i}\right)$. Спроецируем орты пространства свойств на эту плоскость, и будем рассматривать длины полученных проекций:

$$\left\| \Pr_{\alpha\vec{x}_1^{\,i} + \beta\vec{x}_2^{\,i}} (e_t) \right\| = {x_{1ti}}^2 + {x_{2ti}}^2 + x_{1ti} \cdot x_{2ti} \cdot \left(\frac{\left(\vec{x}_1^{\,i}, \vec{x}_2^{\,i}\right)}{\left\|\vec{x}_1^{\,i}\right\| \cdot \left\|\vec{x}_2^{\,i}\right\|} \right) \tag{12}$$

как реализации случайной величины ξ. Ограничимся случаем, когда элементы матрицы X не превосходят по модулю 1[3]. При этом будем рассматривать их как множество независимых, равномерно распределенных на отрезке $[0,1]$ случайных величин. В этом случае мы можем найти функцию плотности распределения ξ. На отрезке $\left[0,\ 1 \Big/ \left(\vec{x}_1^{\,i}, \vec{x}_2^{\,i}\right)\right]$ эта функция выглядит следующим образом:

$$f_\xi(x) = \lambda\left(\frac{1}{2} \cdot \sqrt{x} + 1 \right) \cdot \frac{\ln\left| \frac{x}{\left(\vec{x}_1^{\,i}, \vec{x}_2^{\,i}\right)} \right|}{2} \tag{13}$$

где $\lambda = const$, что обеспечивает нам ненулевое математическое ожидание. Так как размерность пространства очень велика, в силу вступает закон больших чисел. Поэтому при увеличении размерности, полученный в результате проецирования всех орт, многоугольник будет стремится к окружности, что позволяет провести рассуждения, аналогичные случаю шара. Таким образом, (11) верна и для знакоопределенных функций значимости.

В наших экспериментах мы будем сравнивать эффективность исследуемых методов с эффективностью частотного анализа, который принадлежит нашему семейству и имеет единичную матрицу X. Поэтому, формулу оценки можно преобразовать к виду

$$\rho(m_1, m_2) = \frac{1}{\Pi \cdot |T|} \sum_{i=0}^{|T|} \left(\Pi - \arccos\left(\frac{\left(\vec{x}_1^{\,i}, \vec{x}_2^{\,i}\right)}{\left\|\vec{x}_1^{\,i}\right\| \cdot \left\|\vec{x}_2^{\,i}\right\|} \right) \right) \left(\frac{\vec{x}_{1ii}}{\left\|\vec{x}_1^{\,i}\right\|} \right) \tag{14}$$

[3] Нетрудно показать, что умножение на константу не влияет на конечную эффективность, поэтому мы можем ограничится рассмотрением такого случая.

В наших рассуждениях мы полагали, что размер общего словаря остается постоянным. Однако в проведенных нами опытах это условие не выполнялось, что влияет на величину аргумента $\arccos$. Поэтому мы усреднили величину словаря:

$$\rho(m_1, m_2) = \frac{1}{\Pi \cdot |T|} \sum_{i=0}^{|T|} \left(\Pi - \arccos\left(\frac{\vec{x}_{1ii} \cdot t_0}{\left\| \vec{x}_1^i \right\| \cdot |T|} \right) \right) \tag{15}$$

где $t_0 = 20000$ -- средний размер словаря, в наших экспериментах. Несомненно, что такое изменение повлияло на конечную производительность. К сожалению, на данном этапе мы не сумели найти преобразование, сохраняющее эффективность оценки (11) для случая различных словарей.

4. Экспериментальное обоснование

Для проверки эффективности предложенной схемы мы построили семейство методов поиска и сравнили показания классического метода с полученными предварительными оценками. На данном этапе мы рассмотрели лишь случай, когда функции значимости оставались постоянными, а менялась только матрица X .

4.1 Построение семейства

На сегодня, достаточно большое количество методов явно или неявно используют так называемые семантические словари. В большинстве случаев это использование сводится к составлению функции сопоставляющей паре термов их тематическую близость. Очевидно, что такое преобразование может быть записано в табличной форме, что позволяет рассмотреть его как матрицу X нашей схемы.

Очевидно, что хранить такую матрицу невозможно, так как она не разрежена и имеет огромную размерность. Поэтому, применяются методы приближения, позволяющие представить ее в менее ресурсоемком виде без потери большого количества информации. В частности применяется схема наилучшего приближения ранга r или *Сингулярное разложение (SVD)* с ``выкидыванием" наименее значимых сингулярных значений. Так как такой подход позволяет существенно улучшить простую статистику попарной встречаемости слов за счет сглаживания помех, он был назван *Латентно Семантический Анализ (LSA)* [12,4,10].

В качестве исследуемого семейства методов мы рассмотрели методы, построенные на основе этого подхода [2]. Подробное описание схемы работы методов, основанных на LSA, выходят за рамки нашей статьи, поэтому здесь мы лишь укажем конечные значения ключевых параметров:

$$\begin{aligned} A &\approx \widetilde{A} = U_{lsa} \Sigma_{lsa} V_{lsa} \\ X &= A \cdot A^T = U_{lsa} \Sigma_{lsa}{}^2 U_{lsa} \\ f_i &= tf(t_k) \end{aligned} \tag{16}$$

где A -- матрица термы на документы, $U_{lsa}, \Sigma_{lsa}, V_{lsa}$ -- результат LSA разложения, $tf(t_k)$ -- оператор извлечения частоты терма t_k . Размерность LSA выбиралась случайно в пределах от 20 до 30. Построив это семейство, мы сравнили эффективность его элементов с эффективностью стандартного частотного анализа -- вырожденным элементом рассматриваемого семейства в смысле $X = E$.

4.2 Экспериментальная база

В качестве базы для наших экспериментов использовалось множество тестовых документов из коллекции FBIS набора TREC-5[6]. Из этого множества была выбрана 1000 документов, которая содержала 500 документов, релевантных 20

запросам, по которым имелись экспертные оценки, и 500 других случайных документов коллекции. Далее мы выполнили 8 различных случайных выборок из этих 1000 документов. 4 выборки по 400 документов и 4 по 500. После этого мы построили LSA разложение каждой выборки.

Таким образом мы получили 8 независимых выборок и соответствующих им матриц X . В результате проведения поиска по указанным запросам, был построен ранжированный список документов, отсортированный по убыванию ранга. Для вычисления интеграла (2) мы воспользовались следующим выражением:

$$e(f) = \frac{1}{|D|} \sum_{i=1}^{|R|} \frac{(p_i + p_{i-1})}{2} \cdot (n_i - n_{i-1}) \tag{17}$$

где p_i -- точность после i-го правильного документа, n_i -- общее количество документов после нахождения i-го правильного документа, R -- множество релевантных документов. Для усреднения полученных значений использовалось среднее арифметическое.

Оценка(15), с учетом специфики построенного семейства преобразовалась в (18).

$$\rho(m_1, m_2) = \frac{1}{\Pi \cdot |T|} \sum_{i=0}^{|T|} \left(\Pi - \arccos\left(\frac{(a_i, a_j) \cdot t_0}{|T| \cdot \sqrt{\sum_{j=1}^{|T|} (a_i, a_j)}} \right) \right) \tag{18}$$

где a_i -- соответствующие строки матрицы $A = \Sigma U$, используемой в качестве разложения матрицы X ($X = A^T A$).

4.3 Результаты

Поучив по указанной в п.3.2 схеме результаты вычислений $\rho(f_i, tf)$ и $\Delta(f_i, tf)$ и используя метод наименьших квадратов, было получено первое приближение зависимости $\rho(f_i, tf)$ от $\Delta(f_i, tf)$. Угловой коэффициент полученной прямой положителен, что, при незначительных отклонениях от этого приближения, позволяет говорить о монотонной зависимости построенной оценки от сравнительной эффективности, что и требовалось показать. Как уже отмечалось в п.3.2 усреднение мощности словаря сказывается на конечной точности нашей оценки, этот факт можно почерпнуть из Рис.1.

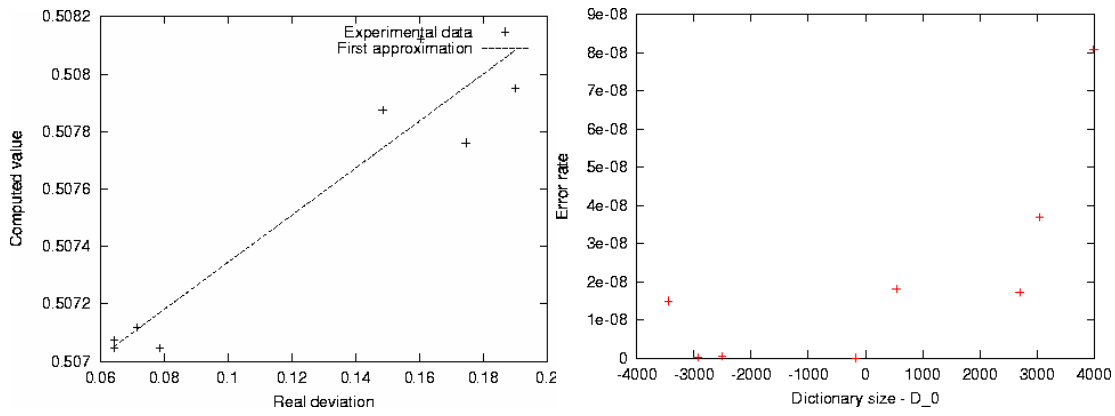

Рисунок 1 Зависимость $\rho(f_i, tf)$ *от* $\Delta(f_i, tf)$ *(справа) и зависимость отклонения от первого приближения от* $|T| - t_0$ *(слева)*

Можно отметить малые отклонения пар оценка/сравнительная эффективность от первого приближения, которая в среднем оказалась равна $1.87 \cdot 10^{-8}$. Этот факт позволяет выдвинуть гипотезу о прямой зависимости этих величин.

Как видно из предположений п. 3.2, кроме оценки сравнительной эффективности мы получили оценку вероятности совпадений результатов сравнений близости векторов свойств. В нашем случае эта оценка оказалась близка к 0.5, что в первую очередь связано со структурой простого частотного анализа, В случае же больших отклонений оценки этой вероятности она также может быть использована для анализа характеристик метода.

Необходимо отметить, что незначительные изменения в построении векторизации значительным образом отразились на эффективности. Более того, оказалось практически невозможно предсказать эффективность конечного метода, анализируя лишь базу его построения.

5. Заключение

В нашей работе мы построили метод предварительной оценки сравнительной эффективности семантических методов. Для оценки точности построенного метода использовалась достаточно большая экспериментальная база, построенная на основе стандартного набора данных и экспертных оценок, предоставляемых Text REtrieval Conference [6]. Несмотря на неплохие экспериментальные результаты, много вопросов остаются открытыми. Среди них такие серьезные как:

Исследование других семейств методов анализа текстов

Более полное исследование *Гипотезы Соответствия*

Рассмотрение других методов разложения матрицы X

Построение предварительных оценок эффективности методов обработки текстовой информации представляется сравнительно новой и мало исследованной областью в IR. Предложенный подход позволит значительно уменьшить временные затраты на исследования в этой области и позволит более четко поставить области построения и оценки эффективности методов обработки текстовой информации.

6. Библиография

[1] Добрынин В. Ю. Кураленок И. Е. "Автоматическая классификация документов". В *Труды Всероссийской научно-методической конференции ``Интернет и современное сообщество''*, Санкт-Петербург", December 1998.

[2] Кураленок И. Е." "Некрестьянов И.С. "Автоматическая классификация документов с использованием семантического анализа". В *Труды первой всероссийской научной конференции ``Электронные библиотеки''*, С-Петербург, Россия, октябрь 1999.

[3] L. Douglas Baker and Andrew Kachites McCallum. Distributional clustering of words for text classification. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Categorisation, pages 96-103, 1998.

[4] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391-407, 1990.

[5] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proc. of the SIGIR'99*, pages 50-57, Berkley, CA, USA, August 1999.

[6] K. Harman. Overview of the Third Text REtrieval Conference. 1997.

[7] L. D. Baker and A. K. McCallum. Distributional clustering of words for text classification. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 96-103, 1998.

[8] Marchiori, Massimo. The Quest for Correct Information on the Web: Hyper Search Engines. In *Proceedings of the Sixth International World Wide Web Conference*, 1997.

[9] Ron Papka and James Allan. Document classification using multiword features. In George Gardarin, James French, Niki Pissinou, Kia Makki, and Luc Bouganim, editors, *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM-98)*, pages 124-131, New York, November 3-7 1998. ACM Press.

[10] P.W.Foltz. Using Latent Semantic Indexing for information filtering. In *ACM Conference on Office Information Systems(COIS)*, pages 40-47, 1990.

[11] G. Salton and J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, NY, 1983.

[12] T. Landauer, P. Foltz, and D. Laham. *Discourse Processes*, volume 25. 1998.

[13] Y. Yang and J. Pederson. Feature selection in statistical learning of text categorization. In *Proc. of the ICML'97*, pages 412-420, 1997.

SEMANTIC METHODS EFFECTIVENESS ESTIMATION IN INFORMATION RETRIEVAL

Igor E. Kuralenok

St.Petersburg State University

In this paper we propose an approach to construction of estimation for information retrieval methods effectiveness without using test collections. We have developed such an estimation for wide class of semantic methods. Our experiments demonstrated close to real efficiency rate results. This fact allows using developed technique as a measure of efficiency if no test collection available.