Система интеллектуального поиска и анализа информации «Exactus» на РОМИП-2009

© Смирнов И.В., Соченков И.В., Тихомиров И. А.

Институт системного анализа РАН
matandra@isa.ru

Аннотация

В статье представлены результаты участия проекта EXACTUS в семинаре РОМИП-2009 по дорожкам поиска и классификации web-страниц. Описан подход к определению информационной значимости терминов документов и его использование в задачах поиска и классификации. Представлены выводы о результатах, полученных с применением предложенного подхода.

1. Введение

В 2009 году система Exactus приняла участие в традиционных дорожках поиска по Web-коллекциям, контекстно-зависимому аннотированию и тематической классификации Web-страниц. В 2009 году были доработаны формулы определения информационной значимости слов естественного языка (ЕЯ) в текстовых документах и модифицирован алгоритм оптимизации параметров поискового алгоритма.

Основной целью участия системы Exactus в РОМИП являлась независимая оценка качества работы алгоритмов. Успешный опыт участия в предыдущих семинарах РОМИП [1], [2] поставил перед авторами вопрос о вкладе семантического анализа текста в улучшение качества ранжирования результатов поиска.

В 2009 году система Exactus дебютировала в дорожке тематической классификации Web-страниц – авторами разработан метод классификации гипертекстовых документов, базирующийся на тех же принципах определения информационной значимости слов ЕЯ, что и статистическая составляющая алгоритма поиска Exactus. В ходе РОМИП-2009 испытаны 2 модификации этого

метода с учётом специфики дорожки тематической классификации Web-страниц.

2. Определение значимости терминов документов в алгоритме семантического поиска Exactus

Неоднократно отмечалось, что поисковый алгоритм Exactus объединяет статистические и лингвистические методы поиска [1], [2], [5]. Тексты ЕЯ представляют собой набор предложений, каждое из которых есть некоторое высказывание. При семантическом анализе множество синтаксем каждого предложения отображается в неоднородную семантическую сеть [8]. В вершинах сети находятся синтаксемы с приписанными значениями (атрибутами), а семантические связи на множестве синтаксем представлены дугами сети. Задача определения смысловой близости состоит в сравнении неоднородных семантических сетей, соответствующих запросу и предложению текста документа, и выборе предложений, наиболее «схожих» с образом запроса.

Метод ранжирования текстовых документов в системе Exactus формализуется следующим образом. Пусть $D(t) = \{S\}$ – множество предложений текстового документа t. Релевантность документа определяется релевантностью «самого лучшего» предложения: $R(t) = \max_{S \in D(t)} \{R(S)\}$. Каждое предложение $S = \{w\}$ – множество вхождений лексем. Запрос пользователя также является предложением: $Q = \{w^q\}$. Через $\tau(t)$ обозначим множество вхождений лексем в текст документа t.

При оценке близости запроса и предложений документа учитывается статистическая релевантность R_W («по ключевым словам») и релевантность по значениям синтаксем R_R:

$$R(S) = \alpha \cdot R_W(S) + (1-\alpha) \cdot R_R(S),\ 0 \le \alpha \le 1.$$

Для оценки соответствия запроса и предложения документа по ключевым словам в предложении S для каждой лексемы w известна парадигматическая форма вхождения $f(w,S)$, при несовпадении формы соответственных лексем в запросе и в

предложении документа вхождение лексемы учитывается с пониженным весом:

$$p_W(w,Q,S)=\begin{cases}1, & f(w,Q)=f(w,S),\\ \beta, & f(w,Q)\neq f(w,S)\end{cases}$$ – где

$0<\beta<1.$

Таким образом, оценка соответствия предложения документа запросу выражается формулой:

$$R_W(t,S,Q)=\frac{|S\cap Q|}{|Q|}\cdot\sum_{w\in S\cap Q}p_W(w,Q,S)\cdot TF\left(w,\tau(t)\right)\cdot IDF(w,C). \quad (1)$$

$\frac{|S\cap Q|}{|Q|}$ – усредняющий множитель. Нормированная на 1 величина инверсной частоты встречаемости слова w в коллекции документов $C=\{t\}$ определяется формулой (2):

$$IDF(w,C)=\frac{\log_2\left(\frac{|C|+1}{|\{t\in C\mid w\in\tau(t)\}|+1}\right)}{\log_2\left(|C|+1\right)}. \quad (2)$$

$TF\left(w,\tau(t)\right)$ – относительная частота встречаемости термина w в тексте $\tau(t)$ [6], [7].

Для оценки семантической близости запроса и предложения документа используется модифицированный вариант формулы (1). При расчёте релевантности сопоставляются значения соответственных синтаксем запроса и предложения: учитываются только те синтаксемы, значения которых совпадают.

Для оценки важности вхождения слова в текст $\tau(t)$ использовались весовые функции следующего вида:

$$A1(w,\tau) = \log_2\left(1 + \sqrt[8]{TF(w,\tau)}\right), \qquad (3)$$

– эмпирическая формула;

$$E(w,\tau) = \gamma \cdot TF(w,\tau) \cdot \log_2\left(\frac{1}{\gamma \cdot TF(w,\tau)}\right), \qquad (4)$$

$$KE(w,\tau) = \sqrt[\eta]{\gamma \cdot TF(w,\tau)} \cdot \log_2\left(\frac{1}{\sqrt[\eta]{\gamma \cdot TF(w,\tau)}}\right). \qquad (5)$$

Здесь $\gamma > 0$, $\eta \in \mathrm{N} \setminus \{0\}$ – параметры.

Формулы (4) – (5) являются частным случаем общей формулы определения информативности слова в тексте документа, предложенной авторами:

$$V(w,\tau) = \varphi\left(TF(w,\tau)\right) \cdot \log_2\left(\frac{1}{\varphi\left(TF(w,\tau)\right)}\right), \qquad (6)$$

где φ – непрерывная функция $0 \le \phi \le 1$ на отрезке $[0;1]$.

Совместно с формулой (2) величины (4) – (6) позволяют аппроксимировать значимость терминов текста с учётом частот встречаемости и могут рассматриваться как альтернатива для подхода к определению информационной значимости слов на основе BM-25 [6], [7].

Алгоритм ранжирования Exactus в модификации 2009 года зависит не более чем от 4 настроечных параметров. Область определения каждого параметра задаётся либо, исходя из определения соответствующих функций, либо эмпирически. Небольшое число параметров алгоритма (α, β, η) позволяет оптимизировать их на основе известных таблиц релевантности путём простого перебора значений параметров из некоторого интервала с заданным шагом. Подобная оптимизация позволяет найти локальные максимумы ранжирующей функции алгоритма.

3. Результаты участия EXACTUS в дорожках поиска РОМИП-2009

В 2008 году были проверены алгоритмы ранжирования с использованием формулы (3) – A1 [1]. В 2009 году выполнено сравнение двух алгоритмов ранжирования результатов поиска: на основе A1 с учётом семантической релевантности и без учёта семантической релевантности – A1-nr. Также был выполнен прогон на основе E-формулы (5).

Графики TREC (OR-оценка) по коллекции BY приведёны на рисунке 1 (прогоны Exactus изображены жирными точечными и штрихпунктирными линиями).

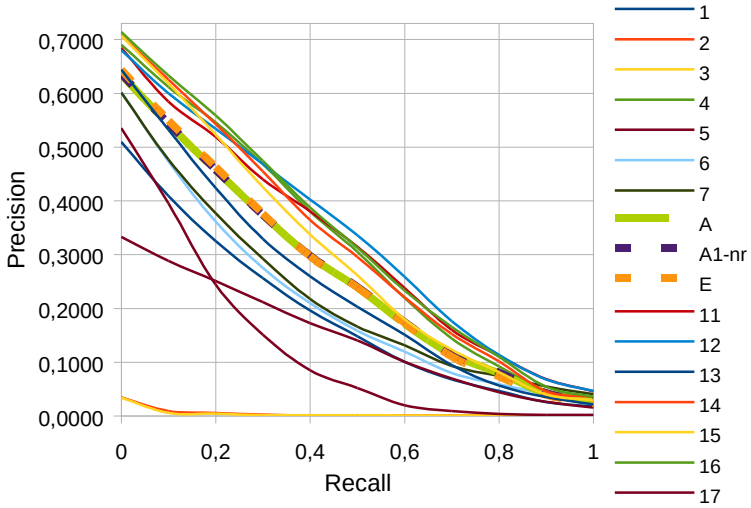

Рисунок1. Графики TREC – OR-оценка для коллекции BY

Из графиков видно, что результаты трёх прогонов системы Exactus отличаются незначительно. При этом алгоритм на основе статистической формулы E (4) примерно совпадает с алгоритмами на основе эмпирических формул A1 (3). Параметры обоих вариантов алгоритма (A1 и A1-nr) были оптимизированы независимо друг от друга: в ходе оптимизации алгоритмов были найдены наборы параметров, соответствующие различным локальным максимумам. Этим объясняется близость значений оценок качества ранжирования.

В силу того, что при расчете значений метрик признаются релевантными все документы, оцененные асессорами хотя бы как relevant-minus, на результатах участников слабо отражается порядок следования документов, признанных релевантными. При этом на метрики в большей степени влияет порядок вхождения нерелевантных документов в выдачу системы (а также их количество). Таким образом, несмотря на то, что относительный порядок результатов в выдаче системе отличался для разных прогонов, полученные результаты оценок не позволяют уверенно говорить о качественных отличиях алгоритмов A1, A1-nr и E.

Вышеприведённые факторы не позволяют достоверно определить превосходство какого-либо метода (A1, A1-nr, E) над остальными и не даёт ответа на вопрос о вкладе семантической составляющей алгоритма в общую оценку релевантности.

С точки зрения алгоритма ранжирования поисковые запросы разделяются на два типа:

1) запросы, в которых определены значения синтаксем;
2) запросы, в которых значения синтаксем не определены.

Запросы первого типа составляют 38,4% от общего числа, в среднем их длина – 4.34 слова. Запросы второго типа составляют, соответственно, 61,6% и имеют среднюю длину 2.46 слова. Более длинные запросы, в которых определены значения синтаксем, содержат более точную формулировку предмета поиска – описание некоторой ситуации. Например, запросы 1-го типа: «*какая из термоядерных реакций наиболее важна для подержания светимости солнца*», «*Природа, сущность и общественная роль политологии*». Запросы 2-го типа содержат более общую формулировку цели поиска: «*москва*», «*праздники*», «*золотая звезда*», «*антенна кавказ продажа*», «*Кижи*».

Алгоритм ранжирования запросов 1-го типа включает расчёт релевантности по значениям синтаксем, тогда как для запросов 2-го типа рассчитывается только статистическая релевантность.

В 2009 году при настройке алгоритма запросы не разделялись на 2 типа, что привело к нахождению локального максимума ранжирующей функции по «усреднённому запросу». Это привело к ухудшению результатов по причине отрицательной корреляции некоторых параметров алгоритма.

При обучении алгоритма ранжирования также не учитывалась степень релевантности обучающего примера запросу пользователя: аналогично методике РОМИП документ признавался релевантным запросу, если он оценен хотя бы одним асессором не

менее relevant-minus. Несмотря на то, что в 2008 году такой подход успешно проявил себя, в 2009 году увеличенная обучающая выборка показала худшие результаты. Таким образом, большая обучающая выборка не всегда способствует повышению качества поиска при применении недифференцированного подхода к обучению.

Вышесказанное относится к результатам участия поискового алгоритма Exactus в дорожке поиска по коллекции КМ. Были испытаны 2 алгоритма ранжирования: на основе формул E (4) и KE(5) (оба с учётом семантической релевантности) – рисунок 2. В отличие от РОМИП-2008 в 2009 году для коллекции КМ была выполнена настройка параметров алгоритма ранжирования на основе таблиц релевантности.

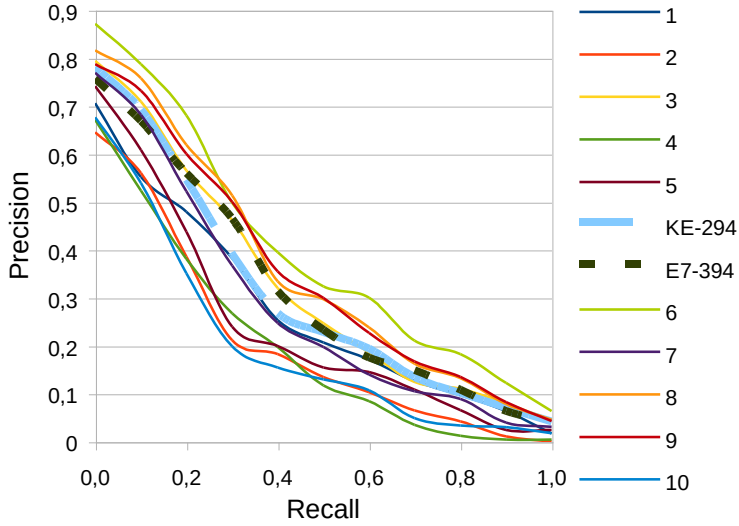

Рисунок 2. График TREC – OR-оценка для коллекции КМ

Из приведённых диаграмм видно, что оба алгоритма показывают стабильные результаты. Алгоритм KE отличается от E лучшей полнотой и точностью на уровне 5, однако проигрывает в средней части графика TREC.

4. Метод автоматической классификации Web-страниц и его результаты на РОМИП-2009

В 2009 году в дорожке автоматической классификации Web-страниц принимали участие 2 метода, базирующихся на тематической IDF и характеристике тематической значимости (ХТЗ) [3]. ХТЗ – есть мера тематической близости между текстом и произвольным классом. ХТЗ базируется на понятии изменения информативности терминов ЕЯ (слов и устойчивых словосочетаний) при отнесении текста к некоторому тематическому классу. ХТЗ представляет собой линейную (по вхождениям терминов в документ) функцию. Решающее правило классификации состоит в сравнении значения ХТЗ с некоторым пороговым значением, которое экспериментально подбирается на этапе обучения классификатора. Для определения информативности вхождений терминов в текст используются формулы (5) – E1 и (6) – E2, соответственно.

На рисунках 3, 4 представлены результаты участников дорожки классификации Web-страниц (OR-оценка).

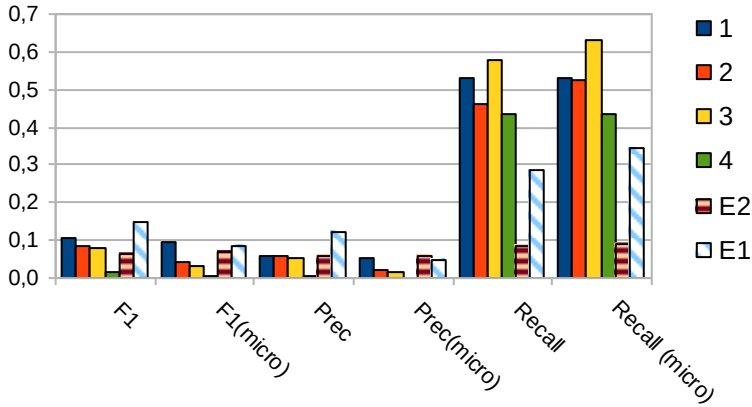

Рисунок 3. Результаты участников дорожки классификации Web-страниц (OR-оценка, с учётом всех возвращённых документов)

Прогоны E1 и E2 отличаются способом подбора значений параметров, используемых в решающем правиле классификации. Для E1 значения параметров были выбрано таким образом, чтобы повысить точность классификации. Эта задача была решена

успешно: метод лидирует по точности (микро- и макроусредение) относительно других участников.

Кроме того, несмотря на меньшую полноту классификации, значение F1-меры также можно считать хорошим: лидирование для прогона E1 по макроусредению.

Для прогона E2 было установлено искусственное ограничение числа возвращаемых документов для каждого класса, и на оценку были переданы ТОП-200 классифицированных документов. Это объясняет относительно низкую полноту классификации. В силу того, что полнота и точность являются взаимосвязанными величинами, полученные результаты на множестве метрик с учётом только оцененных документов уступают результатам других участников.

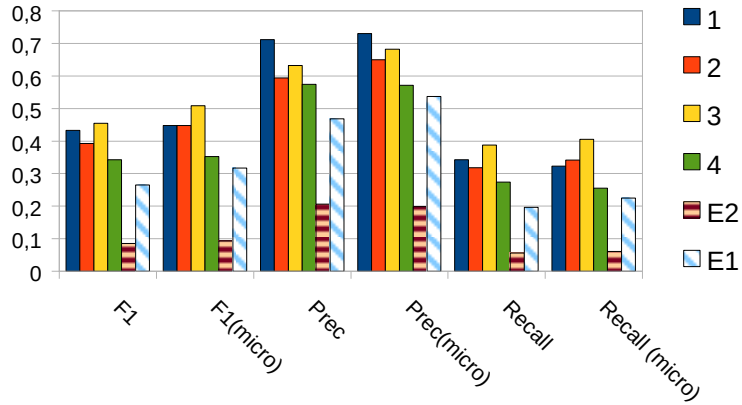

Рисунок 4. Результаты участников дорожки классификации Web-страниц (OR-оценка, с учётом только оцененных документов)

Результаты участия в РОМИП-2009 подтвердили практическую применимость предложенных методов определения информационной значимости, а также тематической близости «текст-класс» на основе ХТЗ.

5. Заключение.

Результаты участия в РОМИП-2009 подтвердили применимость предложенных методов оценки информативности слов в алгоритме ранжирования Exactus. Перспективным представляется раздельная настройка статистической и

семантической ветвей алгоритма в соответствии с принятой в системе классификацией запросов. Необходимо совершенствовать процедуру настройки поискового алгоритма с учётом дифференцированной шкалы оценок релевантности для документов обучающей выборки.

Опыт участия в дорожке классификации Web-страниц позволил выделить дальнейшие направления исследований. Необходима доработка механизма оптимизации параметров, используемых в решающем правиле классификации, с целью нахождения оптимальных соотношений точности и полноты. Дальнейшее развитие метода предполагает извлечение многословных терминов и их учёт в ХТЗ наравне с отдельными словами ЕЯ. Это направление исследований является приоритетным. ХТЗ будет модифицирована для учёта значений синтаксем текста, что должно способствовать повышению качества классификации текстов за счёт определения их смысловой близости.

Литература

[1] Смирнов И.В., Соченков И.В., Муравьев В.В., Тихомиров И. А. Результаты и перспективы поискового алгоритма Exactus. // Труды российского семинара по оценке методов информационного поиска РОМИП'2007-2008. Санкт-Петербург: НУ ЦСИ, 2008, с. 66-76.

[2] Тихомиров И. А. Вопросно-ответный поиск в интеллектуальной поисковой системе Exactus. // Труды четвертого российского семинара по оценке методов информационного поиска РОМИП'2006. Санкт-Петербург: НУ ЦСИ, 2006. - с. 80-85.

[3] Тихомиров И.А., Соченков И.В. Метод динамической контентной фильтрации сетевого трафика на основе анализа текстов на естественном языке. // Вестник НГУ, Информационные технологии, т. 6, Вып. 2, Новосибирск, 2008, с. 94-100.

[4] Золотова Г.А., Онипенко Н. К., Сидорова М. Ю. Коммуникативная грамматика русского языка. Институт русского языка РАН им. В. В. Виноградова, М. 2004 – 544 с.

[5] Osipov G. S., Smirnov I. V., Tikhomirov I. A., Vybornova O.V, Zavjalova O. S. Linguistic Knowledge for Search Relevance Improvement.// Papers of Joint conference on knowledge-based software engineering JCKBSE'06, IOS Press, 2006. - P. 294-302.

[6] Robertson, S. E. Probabilistic models of indexing and searching. / In R.N. Oddy, S.E. Robertson, C.J. van Rijsbergen, P.W. Williams,

Information Retrieval Research, pages 35-56, London, 1981. Butterworths. [Электронный ресурс] http://www.soi.city.ac.uk/~ser/papers/Robertson_vanRijsbergen_Porter.pdf.Проверено 22.05.2009.

[7] Amati, G. Probabilistic models of information retrieval based on measuring the divergence from randomness / G. Amati and C. J. Van Rijsbergen, The Information Retrieval Group, 20(4):357-389, 2002. [Электронный ресурс] http://ir.dcs.gla.ac.uk/terrier/publications/p357-amati.pdf. Проверено 22.05.2009.

[8] Тихомиров И.А. Представление текста в задачах семантического поиска // Сборник трудов 4-го российско-украинского научного семинара "Интеллектуальный анализ информации", Киев 2004,. с. 200-209.

[9] Осипов Г.С., Тихомиров И.А., Смирнов И.В. Exactus – система интеллектуального метапоиска в сети Интернет. // Труды десятой национальной конференции по искусственному унтеллекту с международным участием КИИ-2006. М: Физматлит, 2006. т. 3. - С. 859-866.

Intelligent search engine Exactus:

ROMIP-2009 experience.

© Smirnov I.V., Sochenkov I.V., Tikhomirov I.A.

Institute for Systems Analysis
of the Russian Academy of Sciences
matandra@isa.ru

Abstract

The paper describes EXACTUS results on ROMIP-2009 in ad hoc search and classification tracks. The approach for evaluation of informational importance of documents terms is presented. The results of the presented approach in search and classification tasks are discussed.