Результаты и перспективы поискового алгоритма Exactus

© Смирнов И.В., Соченков И.В.,

Муравьев В. В., Тихомиров И. А.

Институт системного анализа РАН
matandra@isa.ru

Аннотация

В статье описаны усовершенствования поискового алгоритма Exactus, которые позволили в 2008 году значительно улучшить результаты по точности и полноте поиска по сравнению с 2007 годом. Приведены результаты оценки РОМИП'2008 по дорожкам поиска в коллекциях BY, KM и Legal, а также дорожке контекстно-зависимого аннотирования. Также представлен анализ результатов и сделаны выводы о перспективности поискового алгоритма Exactus.

1. Введение

В 2008 году поисковый алгоритм Exactus претерпел значительные изменения по сравнению с 2007 годом. Основной новинкой явилось включение в алгоритм анализа текстов контекстных правил установления значений минимальных синтактико-семантических единиц текста (синтаксем) [8]. Это позволило значительно улучшить качество семантического анализа и снизить шум при поиске. Кроме того, был проведен ряд технических усовершенствований в области обработки документов и преобразования их к рабочему формату системы, что позволило избежать технических потерь при обработке текстов.

Статистические алгоритмы [7], лежащие в основе механизмов расчёта релевантности, были значительно доработаны для повышения качества оценки близости «запрос – предложение

документа» с учётом модификаций классических TFIDF-алгоритмов.

Реализован механизм автоматической настройки параметров ранжирования результатов поиска на конкретную коллекцию, что также позволило повысить точность и полноту поиска.

Дополнительное применение методов разметки текста с учётом тегового окружения различных его фрагментов позволило точнее определять наиболее значимые фрагменты документов.

2. Главная новинка поискового алгоритма Exactus

В различных публикациях неоднократно отмечалось, что поисковый алгоритм Exactus объединяет статистические и лингвистические методы поиска [4,5,7,9]. Лингвистическая составляющая алгоритма заключается в учете смысловых значений слов, которые определяются на основании теории коммуникативной грамматики русского языка [1] с использованием понятия синтаксема. Опишем кратко подход к семантическому анализу текстов, применяемый в Exactus.

Семантический анализ текста имеет своей целью извлечение смысла из текста и отображение его в формальную модель, которая позволяет находить смысловую близость двух текстов. Применительно к задаче поиска – близость запроса и документа. При компьютерном семантическом анализе текста множество синтаксем каждого предложения отображается в неоднородную семантическую сеть [2] с синтаксемами в вершинах и семантическими связями на множестве синтаксем в качестве ребер.

Семантический анализ текста оперирует в основном именными синтаксемами. Именная синтаксема представляется в тексте именной или предложной группой – словосочетанием с существительным или предлогом в качестве управляющего слова. Именная синтаксема характеризуется морфологической формой – предлогом, падежом, и категориально-семантическим классом существительного, от которого она образована. Морфологическая форма синтаксемы и категориально-семантический класс определяются с помощью лингвистического анализатора текста. Синтаксема характеризуется также синтаксической функцией, которую она может выполнять в предложении, и синтаксическим значением. В ходе семантического анализа текста необходимо

установить ***значения*** именных синтаксем, которые являются обозначениями смыслов, передаваемых текстом.

Морфологическая форма и категориально-семантический класс именной синтаксемы не однозначно задают её значение, а синтаксическую функцию, в которой выступает конкретная синтаксема, встречаемая в тексте в ходе анализа, автоматически определить невозможно. Поэтому обычно в анализ вовлекается контекст - глагол или отглагольное существительное, т.е. предикатное слово, при котором именная синтаксема встречается в предложении. Учет такого рода контекста требует создания специального словаря, описывающего наиболее частые сочетания определенного глагола с возможными синтаксемами при нем, и такой словарь был создан для глаголов и отглагольных существительных, наиболее часто встречаемых в текстах определенной тематики.

Словарь предикатных слов не может охватить все глаголы и отглагольные существительные, т.к. перечисление возможных синтаксем при глаголе является весьма трудоёмкой задачей, требующих больших затрат сил лингвистов. Поэтому часто при семантическом анализе невозможно опираться на предикатное слово, так как его нет в словаре предикатов, следовательно, для установления значения синтаксемы в таких случаях необходимо учитывать другой контекст синтаксемы.

В безглагольных предложениях или предложениях, для которых предикатное слово не найдено в словаре, синтаксемы присутствуют рядом с другими элементами предложения, и несут своё значение только в данном контексте. Зависимость значения синтаксемы от собственных морфологических характеристик и характеристик соседних элементов предложения (не глаголов) является языковой закономерностью, которую необходимо обнаружить и зафиксировать для выполнения семантического анализа безглагольных предложений в дальнейшем. Такую закономерность для значений синтаксемы можно записать в виде **правила**, где в посылке правила находятся характеристики самой синтаксемы и окружающих её синтаксем и других элементов предложения, а в заключении правила находится значение, которое необходимо приписать целевой, рассматриваемой синтаксеме [8].

Построение правил установления значений синтаксем экспертом-лингвистом требует больших трудозатрат на просмотр текстов, где встречаются анализируемые синтаксемы, анализ контекста синтаксем, обобщение признаков, влияющих на значение

синтаксемы в разных текстах. Поэтому встала задача автоматического построения **контекстных правил**, позволяющих устанавливать значения синтаксем на основании доступных характеристик самих синтаксем и других элементов предложения, соседствующих с рассматриваемыми синтаксемами.

Материалом для построения обучающих примеров послужила электронная версия синтаксического словаря Г.А. Золотовой, предоставленная сотрудниками Машинного фонда русского языка Института русского языка РАН. В электронной версии словаря границы синтаксем выделены с помощью знаков подчеркивания «_». Это дает возможность автоматически выделить фрагменты текста, содержащие примеры синтаксем, и построить обучающие примеры.

Обучающие примеры – синтаксемы в контекстах строились для всех синтаксем словаря, кроме синтаксем именительного падежа. Для полученного множества обучающих примеров выполнялся метод порождения правил установления значений синтаксем, основанный на ДСМ-методе машинного обучения. Всего было порождено более тысячи правил. Для каждого правила сохранялись примеры, из которых оно было получено. Каждый пример, помимо признаков, содержит тексты, из которых были созданы целевая и соседняя синтаксемы, а также обрабатываемое предложение целиком. Таким образом, каждое правило хранит своё обоснование, которое может быть полезным как для оценивания адекватности реализации метода, так и при анализе лингвистом результатов предсказания на новых примерах.

Для более доступного восприятия была реализована специальная процедура формирования словесной формулировки правил. Словесная формулировка представляет собой описание правила на естественном языке и предназначена для экспертов-лингвистов, обычно затрудняющихся понимать запись в виде математических формул.

Правила сохраняются в размеченный текстовый файл, из которого их можно впоследствии загрузить и использовать для предсказания значений новых синтаксем.

Приведем пример работы процедуры вывода полной текстовой информации для правила установления значения «дестинатив» (назначение предмета или действия) для синтаксемы родительного падежа с предлогом «для»:

Правило: *Если встречается синтаксема в падеже <родительный> с предлогом <для>, имеющая категориальный класс <личное>, а до неё встречается синтаксема в падеже <именительный>, имеющая категориальный класс <предметное>, то полагается, что первая синтаксема имеет значение <дестинатив - назначение предмета или действия >*

Обоснование:
Пример 1:
ЗНАЧЕНИЕ = дестинатив
ЦЕЛЕВАЯ СИНТАКСЕМА = для тебя; КСК: личное
СОСЕДНЯЯ СИНТАКСЕМА = Все; ПРЕДЛОГ: ;ПАДЕЖ: им.вин.;
КСК: предметное; ПОЗИЦИЯ: до
===КОНТЕКСТ: и песни, и силы - Все для тебя.
Пример 2:
ЗНАЧЕНИЕ = дестинатив
ЦЕЛЕВАЯ СИНТАКСЕМА= для различных рачков; КСК: личное
СОСЕДНЯЯ СИНТАКСЕМА = пища; ПРЕДЛОГ: ;ПАДЕЖ: им.;
КСК: предметное; ПОЗИЦИЯ: до
===КОНТЕКСТ: Эти растения - пища для различных рачков

В примерах поле «КОНТЕКСТ» содержит предложение, из которого был построен пример. Получаемые автоматически словесные формулировки правил имеют тот же вид, что и правила, формулируемые экспертом-лингвистом, что позволяет сравнивать их между собой.

Предложенный алгоритм снятия смысловой многозначности синтаксем на основе контекстных правил позволяет выбрать одно значение для синтаксемы из всех возможных, что уменьшает ошибки семантического анализа текста в среднем в 4 раза.

Реализованные алгоритмы установления значений и снятия семантической многозначности синтаксем на основе порожденных правил установления значений синтаксем внедрены в поисковый алгоритм Exactus, что позволило значительно повысить точность семантического анализа.

3. Краткий анализ результатов Exactus на РОМИП'2008

Главная цель коллектива исследователей и разработчиков Exactus – развитие методов искусственного интеллекта и их внедрение в прикладные области, в частности, в поисковые машины.

В 2008 году Exactus принимал участие в дорожках поиска по коллекциям BY, KM, LEGAL и дорожке контекстно-зависимого аннотирования.

По разосланным оргкомитетом РОМИП результатам, наилучшие оценки достались Exactus в поиске по OR-оценке коллекции BY. На рис. 1 показан график TREC для указанных оценок.

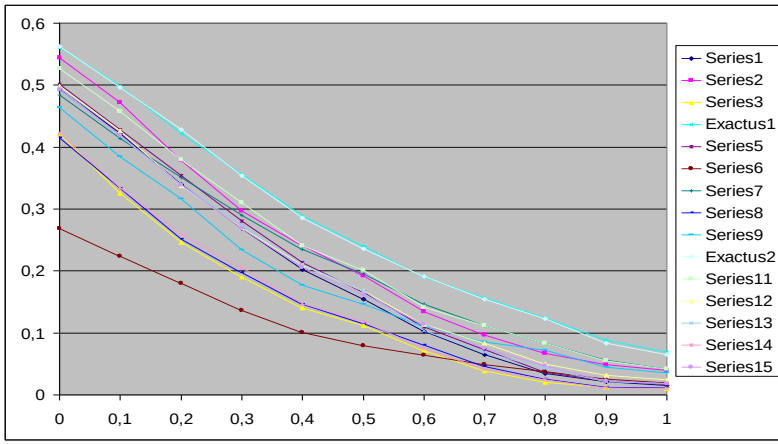

Рисунок 1. График TREC – OR-оценка для коллекции BY.

Разработчиками Exactus были сданы два прогона (пара верхних графиков). Прогоны отличались друг от друга настроечными параметрами поискового алгоритма. По графику видно, что экспериментальный алгоритм Exactus получил ощутимо лучшие оценки по всем точкам TREC-графика и практически по всем оценочным параметрам.

При поиске по коллекции KM также были достигнуты определенные успехи (см. рис. 2).

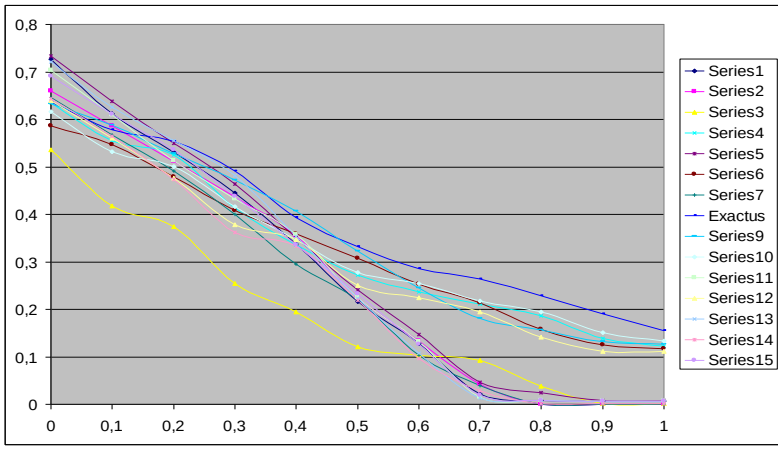

Рисунок 2. График TREC – OR-оценка для коллекции KM.

Как видно из графика, поисковый алгоритм Exactus демонстрирует лучшие результаты в 7 из 11 точках графика TREC, а также по параметрам Bpref, Bpref-10, Recall, Average precision. Незначительный проигрыш по значениям Precision(5) и Precision(10) объясняется успехами других алгоритмов, вероятно, в области оптимизации именно под эти параметры «в жертву» полноте поиска.

Рассмотрим результаты Exactus в поиске по коллекции нормативно-правовых документов (см. рис. 3).

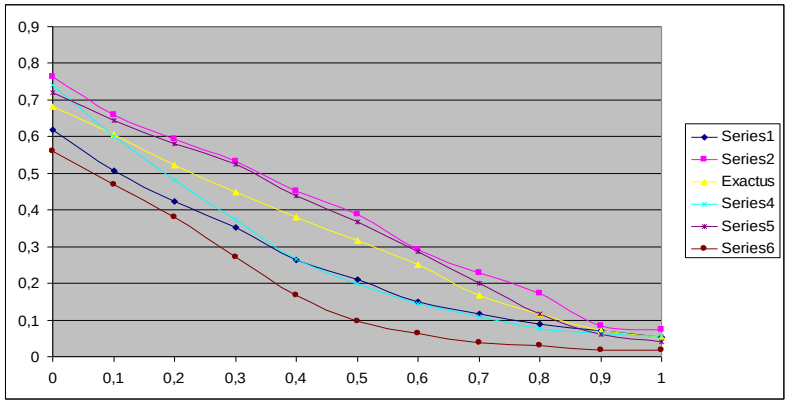

Рисунок 3. График TREC – OR-оценка для коллекции LEGAL.

Поисковый алгоритм Exactus демонстрирует хороший стабильный результат, существенно превосходящий собственные результаты, продемонстрированные в 2007 году. Отставание от лидеров объясняется универсальностью поискового алгоритма Exactus, который не настроен на предметную область и специфическую обработку запросов и нормативно-правовых документов. Так, если в запросе присутствует номер нормативного документа, то нужно начинать поиск именно по этому номеру, занижая веса остальных слов. Аналогичная ситуация с датами – нужно уметь правильно извлекать дату из запроса и находить документы, соответствующие именно этой дате. Кроме того, важным аспектом является фактор времени. Результат не может считаться релевантным, если находится документ с устаревшей информацией, например, приказ, к которому вышло обновление или дополнение.

Таким образом, можно сделать вывод, что для достижения высоких результатов при поиске по этой коллекции необходима тонкая настройка поискового алгоритма на предметную область.

Оценки результатов участия Exactus в дорожке контекстно-зависимого аннотирования приведены на рис. 4.

Из графиков видно, что результаты Exactus в среднем немного уступают другим участникам. Однако из-за малого числа участников в данной дорожке сложно дать качественную оценку полученных результатов. Применённая при оценке данной дорожки шкала также не позволяет с уверенностью оценить, насколько велика погрешность оценок.

Интересным является тот факт, что разброс между min и max оценками по параметру AverageReadability для системы Exactus составляет порядка 0.6. Из этого следует, что мнение асессоров значительно различалось при оценке одних и тех же аннотаций.

В целом результат участия Exactus в дорожке контекстно-зависимого аннотирования является положительным. Однако, для детализированной интерпретации результатов и повышения репрезентативности оценок асессоров необходимо привлечение дополнительных участников и, возможно, пересмотр шкал и параметров оценки качества аннотирования.

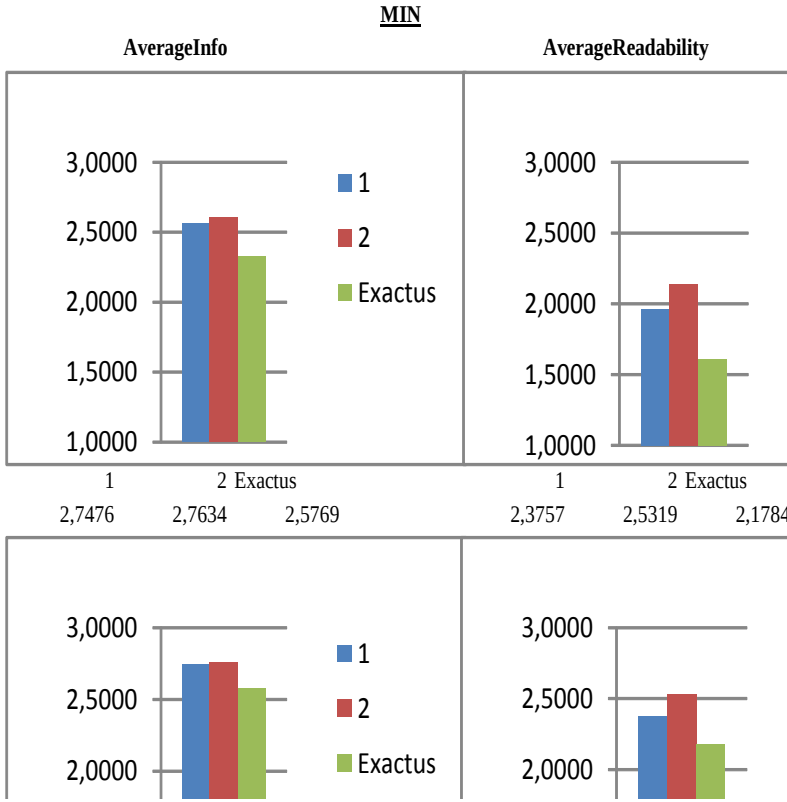

Рисунок 4. Графики оценок информативности и читабельности.

4. Заключение

Полученные в ходе экспериментов РОМИП результаты показывают перспективность экспериментального поискового алгоритма Exactus в сравнении с аналогами. Интересным результатом участия в РОМИП'2008 стало то, что поисковый алгоритм Exactus, не используя алгоритмов ссылочного ранжирования, каталоги сайтов и иные методы рейтинговой оценки веб-документов, показал очень хорошие результаты по точности и полноте поиска, а также по параметру Bpref.

Анализ результатов участия также показал, что нельзя выделить какой-либо один фактор, существенно влияющий на результаты работы поискового алгоритма Exactus. Хороших результатов

позволила достичь совокупность методов, подходов и усовершенствований алгоритмов поиска и анализа текстов, а также хороший уровень технического обеспечения и программирования.

Литература

[1] Золотова Г.А. Синтаксический словарь: Репертуар элементарных единиц русского синтаксиса. – М.: Наука, 1988 – 440 с.

[2] Осипов Г.С. Приобретение знаний интеллектуальными системами: Основы теории и технологии. – М.: Наука, Физматлит, 1997.

[3] 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.

[4] 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.

[5] 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.

[6] 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. 80-85.17.

[7] Тихомиров И.А, Смирнов И.В. Интеграция лингвистических и статистических методов поиска в поисковой машине Exactus //Труды мезденародной конференции Диалог'2008. - С. 485-491.

[8] Смирнов И.В. Метод автоматического установления значений минимальных синтаксических единиц текста. // Информационные технологии и вычислительные системы. – 2008. – №3. – С. 30-45.

[9] Осипов Г.С., Смирнов И.В., Тихомиров И.А. Реляционно-ситуационный метод поиска и анализа текстов и его приложения //Журнал "Искусственный интеллект и принятие решений". Номер 2-2008. - С. 3-10.

[10] Gennady Osipov, Ivan Smirnov, Ilya Tikhomirov, Olga Zavjalova. Application of Linguistic Knowledge to Search Precision Improvement.//Proceedings of 4th International IEEE conference on Intelligent Systems 2008. Volume 2. - P. 17-2 - 17-5.

Results and prospects of Exactus search algorithm.

© Smirnov I.V., Sochenkov I.V.,
Muraviev V.V., Tikhomirov I.A.
Institute for Systems Analysis Russian Academy of Sciences
matandra@isa.ru

Abstract

The paper concerns improvements of Exactus search algorithm which have allowed to refine results on recall and precision in comparison with 2007. The results of ROMIP' 2008 estimation for tracks of adhoc search in collections BY, KM and Legal, and for context-depended annotation are presented. The analysis of results is presented and conclusions on perspectivity of search algorithm Exactus are made.