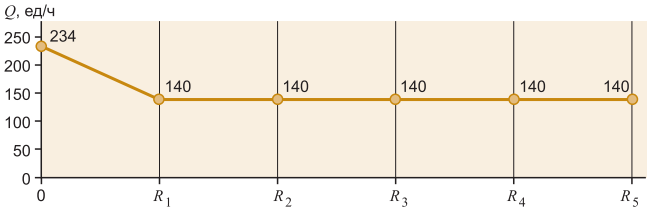

Рис. 7. Зависимость производительности Q сборки от места отказа сборочной позиции R_1....R_5

операциями. Модель сборочной линиикак мультиагентной системы разработана в виде сети Петри. Имитировались частота и место появления отказов, времяустранения отказов. Для обеспеченияотказоустойчивости сборочной линии сравнивались кооперативная и редундантная стратегии обмена программами междупятью роботами,передача программ к началу и концу линии.В результате имитационных экспериментов установлено, что при времени сборочной операции10 с ипоступлениикомплектующих через каждые15 спрограмму сборочнойоперации отказавшего робота следуетпередавать последующему роботу [7].При этомпроизводительность сборочной линииснижается на 40 %независимо от места отказа (рис. 7).

В[8]описаны примерыимитационного моделирования системы управления подземными машинами с поверхности, планирования отгрузки распределенным потребителям, доставки грузов при поступлении случайных заказов, группирования предприятий по максимумуприбыли, многоступенчатой сборки в толкающем и тянущем режимах, проведения подземных выработок, технологии группового обслуживания, сборки в кольцевойлинии. Опыт показал, что разработка и проверка имитационной модели занимаютоколо 90% времени, а на проведение десятков экспериментов приходится10% времени.

Работа выполнялась по гранту Научного Комитета НАТО OUTR CRG №960628 "Имитация и анимация процессов добычи угля в России" и проекту У0043/995 "Подготовка кадров в области информационных технологий производства для Кузбасса" Федеральной целевой программы"Интеграция науки и высшего образования России на 2002-2006 гг.".

Список литературы

1. *Harel D.* Statecharts: a visual formalism for complex systems // Scientific Computer Programming. July. 1987.
2. *Forrester J.* Industrial dynamics: a major breakthrough for decision makers // Harvard Business Review. 1958, v. 36, №.4.
3. *Соболь И.М.* Численные методы Монте-Карло. М.: Наука. 1973.
4. *Banks J., Carson J.S., Nelson B.L., Nicol D.M.* Discrete-Event System Simulation / Prentice Hall. 2000.
5. *Конюх В.Л., Михайлишин А.Ю.* Имитатор сетей Петри и опыт его применения // 2-я Всероссийская науч.-практ. конф. "Имитационное моделирование. Теория и практика". СПб: ФГУП ЦНИИ ТС. 2005. Т.1.
6. *Elliott, M.* Buyer's Guide Simulation // IEE Solutions. May. 2000.
7. *Конюх В.Л., Игнатьев Я.Б.* Обеспечение работы роботизированной сборочной линии при отказе робота // Сборка в машиностроении, приборостроении. 2003. №9.
8. *Конюх В.Л., Зиновьев В.В.* Примеры имитации и анимации дискретных систем // Кемеровский научный центр Сибирского отделения РАН. 2003.

Конюх Владимир Леонидович *– д-р техн. наук, проф. Новосибирского государственного технического университета.*
E-mail: automation2004@list.ru

Автоматизированный поиск семантически близких слов на примере авиационной терминологии[1]

А.А. Крижановский (СПИИ РАН)

Рассмотрен алгоритм поиска семантически близких слов. Оценена точность поиска авиационных терминов. Связь между словами наглядно представлена с помощью визуализации в поисковой программе.

Введение

Специалистам разных отраслей промышленности и направлений науки для совместной работы необходимо находить общий язык. Одна из трудностей в том, что используется разная терминологию (специальная, техническая, жаргонная), разный лексикон. На помощь приходят тезаурусы, то есть специализированные словари, в которых указаны семантические отношения между понятиями, например такие, как: синонимия (*пилот – летчик*), часть-целое (*фюзеляж – самолет*), обобщение-конкретизация (*аэроплан – биплан*).

Таким образом, семантически близкие слова (СБС), рассматриваемые в данной работе, можно определить либо через понятия, связанные семантическими отношениями, либо как слова близкие по значению, встречающиеся в одном контексте. Рассмотрим особенности поиска СБС и причины выбора в качестве источника данных – *вики* (wiki).

Вики – это современный формат гипертекстовой среды, содержащий и ключевые слова, и категории. Вики является одним из примеров Web 2.0, а сайты второго поколения Internet характеризуются тем, что к их созданию привлечены обычные пользователи. На практике, вики – это Web-сайт для совместной работы, где каждый может принять участие в правке статей.

Тематическая направленность каждого вики-документа определена экспертом с помощью катего-

[1]Работа выполнена при финансовой поддержке РФФИ (проекты № 05-01-00151 и 06-07-89242), Президиума РАН (проект № 2.35) и ОИТВС РАН (проект № 1.9)

рий. Эксперт выбирает категории из заданного набора и присваивает их тексту. Например, энциклопедическая статья *"Беспилотный летательный аппарат"* имеет категорию *Авиация*, а статье *"Реактивный ранец"* присвоена категория *Воздушные суда*. Благодаря наличию у текстов тематических категорий можно выполнять целенаправленный поиск слов в определенной проблемной области. В качестве примера была выбрана авиационная тематика.

Одним из наиболее успешных вики-проектов считается Википедия (ru.wikipedia.org), коллективная Internet энциклопедия, содержащая 1,8 млн. документов. Именно эти документы предлагается использовать в качестве корпуса текстов, на основе которых будет выполняться поиск. Текстовые вики-ресурсы были выбраны из-за наличия общего стандарта документов, определяющего единообразные метаданные (заголовок документа, тематические категории); классификации документов.

Современные алгоритмы поиска СБС (например, алгоритмы SimRank [1], Similarity Flooding [2]) не учитывают такую информацию вики-документов, как ключевые слова и категории. Большое число новых документов представлено в формате вики, поэтому необходим поисковый алгоритм, принимающий во внимание особенности описания и структуры вики-документов.

Поиск СБС является подзадачей таких актуальных задач информационного поиска, как: (*i*) расширение / переформулировка запросов с помощью тезаурусов (в поисковых системах), (*ii*) распознавание запроса в запросно-ответных системах, (*iii*) определение значения многозначного слова и (*iv*) автоматическое создание проблемно-ориентированных тезаурусов. Достоинство тезаурусов, построенных с помощью Википедии [3] – это низкая стоимость, постоянное расширение, то есть адекватность современному лексикону, и многоязыковая поддержка.

В данной работе представлены подход и реализация поиска СБС на основе рейтинга вики-текстов в проблемно-ориентированном корпусе с гиперссылками и категориями.

Состояние дел в области поиска СБС

Поиск семантически близких слов связан с теорией графов, а именно с анализом Web-ссылок и поиском на основе данных тезауруса. Поиск СБС с помощью анализа Web-ссылок основан на следующей предпосылке: *отдельной вершине графа соответствует одна Internet-страница*. При этом отдельной Internet-странице может соответствовать понятие[2] либо словоформа[3]. Принятие этой предпосылки позволяет перейти к задаче поиска *похожих Internet-страниц*, связанной с задачей вычисления меры сходства между вершинами графа. Существует ряд алгоритмов, предлагающих решение этих задач (HITS, PageRank, ArcRank и др).

Сложность организации поиска СБС (и оценки качества поиска) определяется рядом причин. Во-первых, понятие семантической близости определено не для слов, а для значений слов, то есть неразрывно связана с контекстом. Во-вторых, язык – это вечноизменяемая субстанция. Слова могут устаревать или получать новые значения. Особенно активное словообразование и присвоение новых значений словам наблюдается в науке, в ее молодых, активно развивающихся направлениях. В-третьих, нет однозначного общепринятого способа вычисления близости значений слов. Это обуславливает сложность численной оценки работы алгоритмов автоматического поиска СБС.

Автоматический поиск синонимов и СБС является одной из задач автоматической обработки текста. Проведенный анализ проблемы автоматизированного построения списков СБС показывает, что здесь можно выделить следующие основные компоненты:

1. *разработка (выбор существующих) алгоритмов*. Требованием к алгоритму (для поиска семантически близких слов) является учет тех дополнительных возможностей, которые предоставляет рассматриваемый корпус документов. Это наличие категорий (классифицирующих документы по их тематической принадлежности) и метаинформации в виде ключевых слов (например, заголовок документа);

2. *оценка результатов работы алгоритма*. Анализ работ в данной области показывает, что необходима разработка оригинальных показателей степени синонимичности полученных списков семантически близких слов. Задача выбора текстового ресурса определяет способ оценки результатов поиска. Среди множества проблем создания, использования корпусов можно выделить общую проблему отсутствия единого стандарта. В работе в качестве корпуса вики-текстов предлагается использовать коллективную он-лайн энциклопедию Википедия из-за большого числа документов (1,8 млн. ед.) и открытого доступа к ним;

3. *программные ресурсы для обработки текста*. При выборе программных инструментальных средств разработки и проектировании архитектуры программного комплекса, автор придерживался следующих требований: открытость исходного кода, кроссплатформенность, модульность архитектуры, использование общепринятых стандартов, визуализация результатов работы;

4. *визуализация результатов поиска*. Анализ поисковых систем показывает, что некоторые из них обеспечивают визуализацию результатов поиска в виде статического и динамического изображения.

[2] В Википедии (ru.wikipedia.org): странице энциклопедии соответствует некоторое понятие, которое раскрывается в данной энциклопедической статье.

[3] В Викисловаре (ru.wiktionary.org): страница словаря описывает одну словоформу, которая может содержать несколько значений.

Алгоритм

Требуется решить задачи: поиска семантически близких слов, оценки результатов поиска. Входными данными для поиска семантически близких слов являются исходное слово, корпус документов и список слов, уточненный пользователем[4].

Из двух классических алгоритмов HITS и PageRank, удовлетворяющих приведенным требованиям (учет ключевых слов, гиперссылок), был выбран алгоритм HITS по следующим причинам: формулы вычисления в PageRank требуют экспериментального подбора коэффициента; значения весов (рассчитанные с помощью PageRank) не могут быть использованы напрямую для поиска похожих страниц (нужен дополнительный алгоритм, который будет искать похожие страницы на основе весов PageRank).

В HITS алгоритме для поиска Internet-страниц (соответствующих запросу пользователя) предлагается использовать информацию, заложенную в гиперссылки. Демократическая природа Интернет позволяет использовать структуру ссылок как указатель значимости страниц (эта идея присутствует в алгоритме PageRank). Автор страницы p, ссылаясь на страницу q, указывает на авторитетность q. Для алгоритма поиска существенно, что содержание страницы q соответствует тематике страницы p.

HITS алгоритм [4] использует такие понятия, как: авторитетный документ и хаб-документ. *Авторитетный документ* – это документ, соответствующий запросу пользователя, имеющий больший удельный вес среди документов данной тематики, то есть большее число документов ссылаются на данный документ. *Хаб-документ* содержит много ссылок на авторитетные документы.

Формальная постановка задачи, которую решает HITS алгоритм: дан орграф $G = (V, E)$, где V – вершины (документы), E – дуги (гиперссылки). Для каждого документа p известны два списка: (*i*) документы, на которые ссылается данный документ, и (*ii*) документы, ссылающиеся на данный документ. Необходимо найти набор релевантных документов (соответствующих запросу), на которые при этом ссылаются многие документы (нужно найти авторитетные документы).

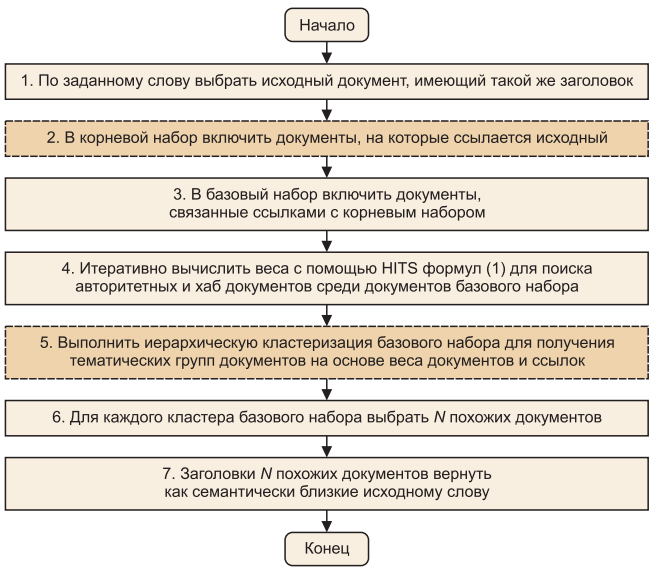

Рис. 1. Адаптированный HITS алгоритм

Каждому документу в HITS алгоритме сопоставляются веса a (*authority*) и h (*hub*), которые показывают, соответственно, насколько документ является авторитетным и насколько он является хорошим хаб-документом. Формулы алгоритма HITS для итеративного вычисления весов таковы:

$$h_j = \sum_{i:(j,i)\in E} a_i; \quad a_j = \sum_{i:(i,j)\in E} h_i,$$

где h_j и a_j показывают соответственно насколько документ j является хорошим указателем на релевантные документы (j-ый документ рассматривается как хаб-документ) и является авторитетным документом.

Адаптированный HITS алгоритм (AHITS), разработанный автором, учитывает метаинформацию проблемно-ориентированного корпуса документов: ключевые слова; категории, классифицирующие документы по их тематической принадлежности; гиперссылки. В HITS алгоритме граф содержит взвешенные вершины одного типа. Предлагается модификация алгоритма для учета трех типов вершин (авторитетный документ, хаб-документ и категория) и трех типов дуг (документ-документ, документ-категория и категория-категория), определяемых проблемно-ориентированным корпусом текстов. Шаги AHITS алгоритма представлены на рис. 1. Шаги, предложенные автором, выделены пунктиром.

Таким образом, исходными данными для алгоритма являются: сеть документов-энциклопедических статей (вершины – документы, дуги – гиперссылки) и дерево категорий (вершины – категории, дуги связывают категории-родителей и детей). Причем элемент сети (статья) связан с одним или несколькими элементами дерева (категории). Для каждого документа определены: (*i*) список документов, ссылающихся на данный документ, (*ii*) список документов, на которые ссылается данный документ, (*iii*) список категорий, определяющих его тематическую принадлежность.

Реализация

Данный алгоритм был реализован в программном комплексе *Synarcher*[5].

Достоинствами поискового комплекса являются: визуализация результатов поиска; уточнение запроса в ходе работы пользователя с программой.

В архитектуру программного комплекса *Synarcher* (рис. 2) включены модули: *kleinberg*, предоставляющий доступ к данным Википедии и реализующий ал-

[4] Список слов, найденный системой и уточненный пользователем, то есть предполагается обратная связь для уточнения результатов поиска.

[5] Программная реализация: *http://synarcher.sourceforge.net*

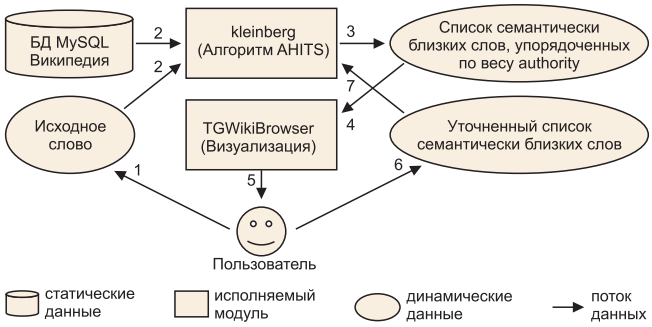

Рис. 2. Архитектура программного комплекса Synarcher

Таблица 1

Синонимы	БПЛА, БЛА
Гиперонимы	летательный аппарат
Гипонимы	спутник, зонд, ракета, автоматическая межпланетная станция
Меронимы	автопилот

горитм AHITS; визуализации *TGWikiBrowser*. Входными данными для AHITS алгоритма являются слово, заданное пользователем, и данные Википедии. Алгоритм строит список СБС упорядоченных по весу authority, а пользователь получает возможность работать с ними благодаря модулю визуализации *TGWikiBrowser*. В ходе работы пользователь уточняет список СБС и запускает алгоритм повторно.

Модуль *kleinberg* программы *Synarcher*. предоставляет доступ к Википедии, хранимой в БД MySQL, размещенной локально или удаленно; позволяет задать параметры AHITS алгоритма; обеспечивает хранение параметров поиска и слов, помеченных пользователем как синонимы на ПК пользователя.

Модуль визуализации написан на основе кода программы визуализации вики-страниц – *TouchGraph WikiBrowser*. Для более удобной навигации код программы был существенно модифицирован, а именно в контекстное меню добавлены команды: спрятать все вершины, пометить вершину как синоним, показать категории.

Графический интерфейс программы состоит из экранов: *Article*, позволяющий просмотреть энциклопедическую статью, соответствующую выбранному слову, *Database*, позволяющий подключиться к БД и получить статистику по БД, Synonyms, где задаются параметры AHITS алгоритма, выводятся результаты поиска в табличной и текстовой форме, и экран с результатами поиска семантически близких слов в виде графа.

Эксперименты

В работе [5] представлены эксперименты по поиску синонимов в английской и русской версии Википедии с помощью AHITS алгоритма и описана сессия поиска синонимов в программе *Synarcher*. Эксперименты показывают, что разработанный программный комплекс *Synarcher* позволяет найти синонимы и семантически близкие слова в английской Википедии, отсутствующие в современных тезаурусах WordNet, Moby (например, найден синоним *Spationaut* для слова *Astronaut*). Тем не менее, некоторые синонимы, представленные в тезаурусах и Википедии, не были найдены.

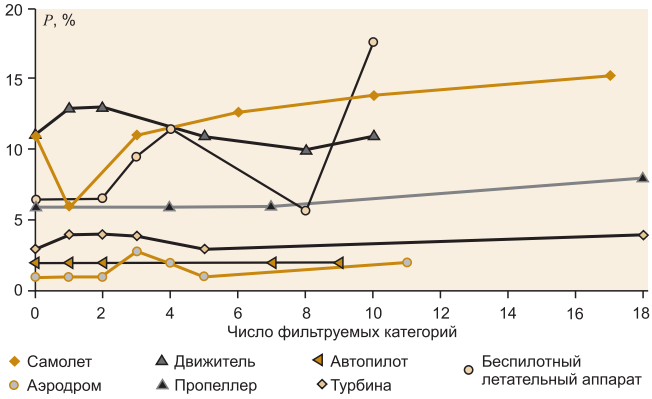

Рис. 3

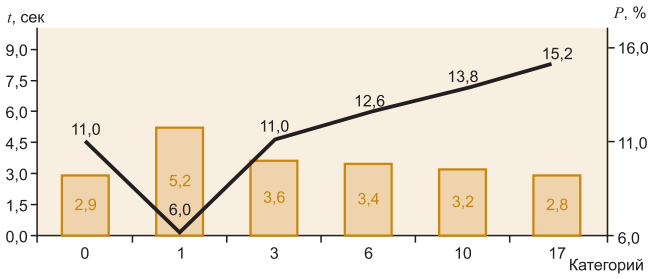

Рис. 4

Точность поиска (*P*), оцениваемая в экспериментах, – это отношение числа семантически близких слов, найденных программой, к общему числу найденных слов. Примеры СБС для словосочетания *"Беспилотный летательный аппарат"* представлены в табл. 1.

Для семи авиационных терминов, для которых есть энциклопедические статьи в русской Википедии (*Автопилот, Аэродром, Беспилотный летательный аппарат, Движитель, Пропеллер, Самолет, Турбина*), проведена серия экспериментов (рис. 3) для оценки времени работы и точности поиска AHITS алгоритма в зависимости от числа категорий. Черный список категорий (blacklist) составляется экспертом и сужает пространство поиска. Например, включение категории "*XX век*" в blacklist позволяет отсечь множество документов с заголовками: *1900, 1901, 1902* и т.д. В эксперименте для фильтрации выбираются категории с максимальным числом слов, не являющихся семантически близкими заданному слову. Для слова *Автопилот* точность поиска была низкой (2%) и не менялась при изменении числа фильтруемых категорий. Возможно, это объясняется недостаточным (по сравнению, например, со статьей *Самолет*) числом ссылок, связывающих статью *Автопилот* с другими. На рис. 4 представлена зависимость точности поиска СБС для слова *Самолет* и времени работы алгоритма от числа фильтруемых категорий, менявшегося от *0* до *17*. Рис. 4 показывает, что при использовании категорий (для слова *Самолет*) время работы вырастает, но точность поиска также увеличивается.

Русской Википедии соответствует орграф, содержащий 171 тыс. вершин, 3.4 млн дуг (на 11.05.07). При поиске в графе AHITS алгоритм строит базовый набор с числом вершин 200...800 ед., числом дуг 800...12 000 ед. (для слова *Самолет*). Указан диапазон вершин и дуг, поскольку изменение фильтруемых категорий меняет число вершин, включаемый в базовый набор. Таким образом, рис. 4 обобщает результаты шести опытов с разными размерами базовых наборов, построенных для слова *Самолет*.

Основная разница HITS и AHITS алгоритмов – не учет/учет категорий соответственно. При числе категорий ноль (первый вертикальный ряд на рис. 4) работа AHITS алгоритма (по скорости и точности поиска) соответствует работе HITS алгоритма. Это позволяет сравнить HITS и AHITS алгоритмы. Сравнение для указанных семи слов (рис. 3) показывает, что работа AHITS алгоритма медленнее HITS алгоритма в среднем на 51%, при этом точность поиска AHITS алгоритма выше на 25%.

Для численной оценки степени сходства эталонного списка и автоматически построенного списка СБС адаптирован коэффициент Спирмена. Адаптация позволяет сравнивать ранжирование элементов в списках разной длины. Итак, для исходного слова даны: эталонный список A, построенный экспертом, и список B, построенный автоматически. В конец списка B добавляются элементы A, в нем отсутствующие. Каждому элементу списка назначается ранг $1...N$. Далее сравниваются положения в списках общих элементов, а именно вычисляется сумма модулей расстояний между i-ми элементами набора (S – число общих элементов):

$$F^S(s_1,s_2) = \sum_{i=1}^{S} |s_1(i) - s_2(i)| .$$

Предложенная модификация коэффициента Спирмена позволила оценить чувствительность результатов AHITS алгоритма к изменению параметров поиска с помощью экспериментов. Для ряда слов из русской Википедии (например, Самолет) точность поиска была достаточно стабильной (значение стандартного отклонения коэффициента Спирмена 4.41), что избавляет пользователя от необходимости тщательно подбирать параметры поиска.

Заключение

Поисковый алгоритм HITS адаптирован к использованию особенностей вики-документов, таких как наличие категоризации, определяющих тематику документа; четкое соответствие заголовка документа и его содержимого, то есть слова заголовка можно рассматривать как ключевые слова документа. Таким образом, реализация нового алгоритма AHITS (адаптированный HITS) позволяет находить документы похожие на заданный. Решение такой задачи сейчас востребовано и некоторые современные поисковые системы позволяют выполнять поиск похожих документов, например *Google*, *Яндекс*.

Программный комплекс *Synarcher*, реализованный на языке программирования *Java*, позволяет выполнять поиск семантически близких слов в энциклопедии Википедия с динамической визуализацией результатов поиска. Визуализация в поисковой программе обеспечивает наглядное изображение связей между словами, что является удобным дополнением к списку найденных слов.

Гиперссылки используются внутри вики-документов для отсылки к документам, описывающим упоминаемые термины. Поэтому учет результата поиска похожих документов поможет в указании недостающих гиперссылок между связанными по смыслу документами, что улучшит связь между документами и повысит качество корпуса вики-документов в целом. Поиск с помощью программного комплекса Synarcher энциклопедических статей Википедии близких по смыслу к заданной статье позволит пользователям более глубоко изучить исследуемое понятие.

Список литературы

1. *Jeh G., Widom J.* SimRank a measure of structural-context similarity. In Proceedings of the Multi-Relational Data Mining Workshop, 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining., 2002.
2. *Melnik S., Garcia-Molina H., Rahm E.* Similarity flooding: a versatile graph matching algorithm and its application to schema matching. In 18th ICDE. San Jose CA, 2002.
3. *Milne D., Medelyan O., Witten I.H.* Mining domain-specific thesauri from Wikipedia: a case study. In Proc. of the International Conference on Web Intelligence (IEEE/WIC/ACM WI'2006). Hong Kong, 2006.
4. *Kleinberg J.* Authoritative sources in a hyperlinked environment. – Journal of the ACM, 1999. Vol. 5, No. 46.
5. *Крижановский А.А.* Автоматизированное построение списков семантически близких слов на основе рейтинга текстов в корпусе с гиперссылками и категориями // Компьютерная лингвистика и интеллектуальные технологии. Тр. междун. конф. "Диалог 2006". Бекасово, 2006.

Крижановский Андрей Анатольевич *– научный сотрудник Санкт-Петербургского института информатики и автоматизации РАН (СПИИРАН).*

Контактный телефон (812) 328-80-71 E-mail: aka@iias.spb.su