УДК 004.9

Система построения понятийной иерархии для ассоциативного поиска по текстам

Н.А. Андреева, П.П. Кокорин

Предложен метод построения понятийных иерархий для ассоциативного поиска по текстам. Приведено описание системы, основанной на предлагаемом методе, а также показана применимость подхода для задач e-Learning.

Сегодня e-Learning (дистанционное обучение) развивается в мире довольно динамично. Причинами такого развития в этой области служат, с одной стороны, увеличение количества компаний, имеющих разветвленную систему филиалов, а с другой — демографический фактор и, как следствие, рост числа студентов во всех странах мира.

С одной стороны, учебным заведениям электронное обучение дает возможность расширения работы с филиалами, привлечения иностранных студентов, снижения затрат на курсы и т.д. С другой же, коммерция все больше нуждается в электронном обучении. Компании, разворачивающие бизнес во многих странах, серьезно выигрывают на e-Learning при обучении персонала в отдаленных филиалах. [http://www.cnews.ru/reviews/index.shtml?2007/09/26/267729].

Однако, организация процесса дистанционного образования (или самообразования), как самоорганизующегося процесса, сопряжена с рядом трудностей, таких как: поиск основополагающих тематически-ориентированных текстов предметной области, определение минимально достаточного понятийно-терминологического базиса, покрывающего всю предметную область и др. Вышеуказанные трудности в большей степени связаны с быстрым увеличением объемов информации, и как следствие этого возросшей сложностью поиска.

Рост информационных ресурсов Интернет-пространства и их потребителей уже достиг критического уровня эффективного функционирования. Увеличение количества информационных ресурсов происходит за счет дублирования, а также примитивной компиляции. Это естественный процесс, развивающийся за счет новых публикаций, а также за счет «журнализации событий» развивающегося мира – новостей. Подобное экстенсивное увеличение количества доступных ресурсов дает большую нагрузку на поисковые машины, которые должны адекватно отвечать на запросы пользователей и выдавать адекватные, актуальные и достоверные данные.

Рост количества бесполезных документов, направленных на поисковый спам, уже превосходит в размерах объем полезных. Рекламные, игровые ресурсы, а также просто каталоги бесполезных ссылок на такие же каталоги «забивают» образовательные и справочные ресурсы, которые должны быть доступны в первую очередь.

Подобная проблема возникла за счет особенности функционирования современных поисковых машин при поиске вхождений интересующих слов в документ без проверки реальной полезности и качества самой страницы.

Возможным решением являются *инфологические информационные системы*. Рассмотрим основные этапы формирования такой информационной системы на примере инфологической системы сайта <www.visualworld.ru>, включающей: информационное представление проблемы, выбор избранных тем (составление антологии), формирование предметных словарей (тезаурусов и глоссариев), формирование онтологии предметной области [1, 2].

Первая фаза заключается в выборе для заданной темы уже опубликованных основополагающих текстов, т.е. в осуществлении антологии. Антология – сборник тематически-ориентированных текстов и является платформой составления тезаурусов и глоссариев.

Глоссарий в рамках семиологической информационной системы – множество терминов, являющееся минимально достаточным понятийно-терминологическим базисом предметной области, в котором все элементы иерархически и ассоциативно связаны с другими терминами заданной предметной области. Связями назовем отношение на множестве терминов, дающее связь между определяемыми и определяющими словами.

Результат же информационного анализа и исследования, представленный как краткий реферат, составляет суть понятия онтология.

При этом актуальной является проблема построения визуальных ассоциаций по заданной предметной области и задача правильного составления антологии.

Необходимо отметить, что термин ассоциации используется достаточно широко в различных значениях. В данном случае *ассоциация* – это иерархия понятий в глоссарии предметной области. С помощью ассоциативных связей возможно переупорядочение понятий онтологии выстраиванием их в определенные понятийные уровни. Для каждого отдельного понятия можно задать его понятийно-ассоциативную иерархию во главе которой он находится. Ассоциированные с ним понятия образуют понятийное окружение для выбранного термина. В то же время само это понятие может входить в другие понятийные окружения. Рекурсивно расширяя данный подход на все понятия онтологии, мы получаем глоссарий предметной области.

Преимущества понятийной визуализации заключаются в том, что пользователь получает инструмент для быстрого ознакомления с нужной предметной областью, не тратя времени на поиск основных понятий и определений, и далее итерационно развивать познание терминов выделенной предметной области.

Рассматриваемая программа, использующая <www.visualworld.ru>, позволяет осуществить самоорганизующиеся процесс Интернет обучения, т.е. e-Learning. На базе поисковой платформы <www.visualworld.ru> реализуется самоорганизующийся процесс обучения понятийной терминологии выделенной предметной области. Преимущество данной поисковой платформы заключается в том, что она не проводит поиск точного совпадения поискового запроса, а пытается найти информацию с максимально точным описанием того, что содержится в запросе. Кроме того, чтобы система смогла лучше «понять», какие нужны данные, по каждому запросу предлагаются связанные с ним ассоциативными связями слова, которые помогают точнее описать требуемую информацию.

Основы семиологического подхода

Коммуникация – установление однозначных связей между любым знаком и символом. В силу специфики возможностей человека с его речевой, текстуальной и визуальной формами восприятия имеются разные формы коммуникативного общения. Интерфейсные возможности компьютера, с одной стороны, развивают коммуникативные возможности, а с другой – жестко ограничены по эффективности. Сегодня наиболее предпочитаемый способ интерфейсного общения между человеком и компьютером, а также между компьютерами – это текстуальная форма.

С точки зрения семиологии, можно подчеркнуть, что однотипность текстуальной формы коммуникативного общения не предполагает требования полной лингвистической совместимости между человеком и компьютером [3].

У. Эко вводит понятие семиологии как основы коммуникативного общения между компьютерными информационными системами. Это поиски вербально-лингвистического механизма для выявления и установления связей между различными словарями, синонимами и семантикой внутри заданного информационного поля.

В рамках семиологии основное внимание уделяется однозначно понимаемой терминологии подобия, ассоциации, однозначности.

Необходимость введения понятия семиологии связано с особой ролью компьютерного коммуникативного общения. Очень многие функции лексики, прагматики и грамматики, изучаемые под понятием «лингвистика», обычно ориентируются на специфические свойства вербального мышления человека. И, следовательно, текстуальный интерфейс при своей компьютерной реализации требует вве-

дения специфических операций, которые условно можно называть семиологическими, включающими в себя такие понятия, как: антология, онтология, идентификация и ассоциация.

Компьютерно-ориентированный процесс идентификации связан с классификацией, ассоциацией, анализом и синтезом текстуального описания и базируется на принципе идентификации неразличимости, который был сформулирован Г. Лейбницем: «Два объекта считаются неразличимыми, если все их свойства общие». Этот принцип наиболее активно используется при индексном ключевом поиске в компьютерных системах (поиск по однозначному совпадению), в котором в качестве ключа выступают буквы, слова, символы и т.д.

В рамках информационных систем можно привести следующее утверждение. Пользователь – это манифестация семантической интерпретации предложения; предметная область – это принцип денотации, а информационная база знаний – принцип сигнификации.

Семиологическая система поддерживается в актуальном состоянии с помощью интерактивного сопровождения пользователем или в автономном режиме путем обновления антологии.

Алгоритм построения понятийной иерархии и тематического глоссария предметной области

Разработанный алгоритм основан на формировании тезауруса и построении глоссария. Для этого предложения в анализируемом тексте разделяются по точкам; все слова приводятся в нормальную форму (им. падеж, ед.ч. для имен существительных, 1 л., ед.ч., наст. время для глаголов и т.д.) с помощью метода морфологического анализа, предложенного в [4]. Метод морфологического анализа работает на основе прикладного морфологического анализа без словаря. Алгоритмы морфологии построены на самообучении программы на открытых массивах реальных текстов и совмещают два подхода: лингвистический – формализованная грамматика для построения морфологических гипотез и математический – метод корреляции, позволяющий унифицировать морфологическую гипотезу. Далее из текста исключаются стоп–слова (малоинформативные слова, используемые в качестве союзов, предлогов, местоимений и т.д.).

Далее формируются связи между всеми словами в каждом предложении. Запускается счетчик, учитывающий, сколько раз та или иная связь слов будет встречаться в тексте (по умолчанию около каждой связи изначально ставится показатель частоты ее упоминания, равный единице). Если связи слов встречаются повторно в следующем предложении, то показания счетчика увеличиваются на единицу. После проведения подсчета частоты упоминания связей исключаются те связи, которые упоминались менее двух раз, и отдельные слова, оставшиеся без связей. В результате формируется список слов, отсортированных по частотам, и список связей. Далее для каждого отдельного слова подсчитывается число связей. На основании всех проведенных операций строится понятийная иерархия – визуальный граф.

По виду визуального графа можно сделать вывод о качестве текста. В идеале в графе не должно быть распавшихся (бессвязных) областей, все слова должны быть связаны друг с другом. Критерием качества построенного графа является отношение количества распавшихся областей к общему количеству слов в документе.

Алгоритм ассоциативного поиска текстов

Формально процесс ассоциативного поиска по текстам можно определить как выборку множества документов по поисковой фразе, удовлетворяющих условию наличия семантических связей в документе между всеми словами поисковой фразы.

Назовем поисковой фразой множество слов, полученных из естественно-языкового запроса, путем приведения всех слов в нормальную форму и удаления слов, являющихся стоп-словами.

Для поисковой фразы производится формирование списка связей между словами, по алгоритму, описанному [3].

Поиск документов, отвечающих поисковой фразе, производится путем выбора документов, удовлетворяющих условию наличия всех связей между словами, имевшимися в поисковой фразе.

Испытания метода ассоциативного поиска текстов

Инфологическая поисковая система <www.visualworld.ru>, реализующая предложенный алгоритм ассоциативного поиска текстов, использовалась в качестве платформы для создания программы поддержки самоорганизующегося итерационного процесса Интернет – обучения.

В качестве примера применения системы, была поставлена задача написать реферат по теме «Нейрокомпьютер». Реферат должен был в достаточной степени охватывать заданную предметную область. В качестве инструментария использовалась описываемая программа и непосредственно сайт <www.visualworld.ru>.

На первом этапе строилась понятийная иерархия термина «нейрокомпьютер» (рисунок). В качестве онтологической базы использовался «Большой энциклопедический Словарь» 1998 года издания. На основе словарной статьи для термина «нейрокомпьютер» был построено его понятийное окружение (на рисунке выделены серым цветом). Текст словарной статьи приведен ниже:

«Нейрокомпьютер – ЭВМ, которая состоит из большого числа параллельно работающих простых вычислительных элементов (нейронов). Элементы связаны между собой, образуя нейронную сеть. Они выполняют единообразные вычислительные действия и не требуют внешнего управления. Большое число параллельно работающих вычислительных элементов обеспечивают высокое быстродействие».

Далее итерационно строятся понятийные окружения терминов второго и последующих уровней иерархии (на рисунке термины на светлом фоне).

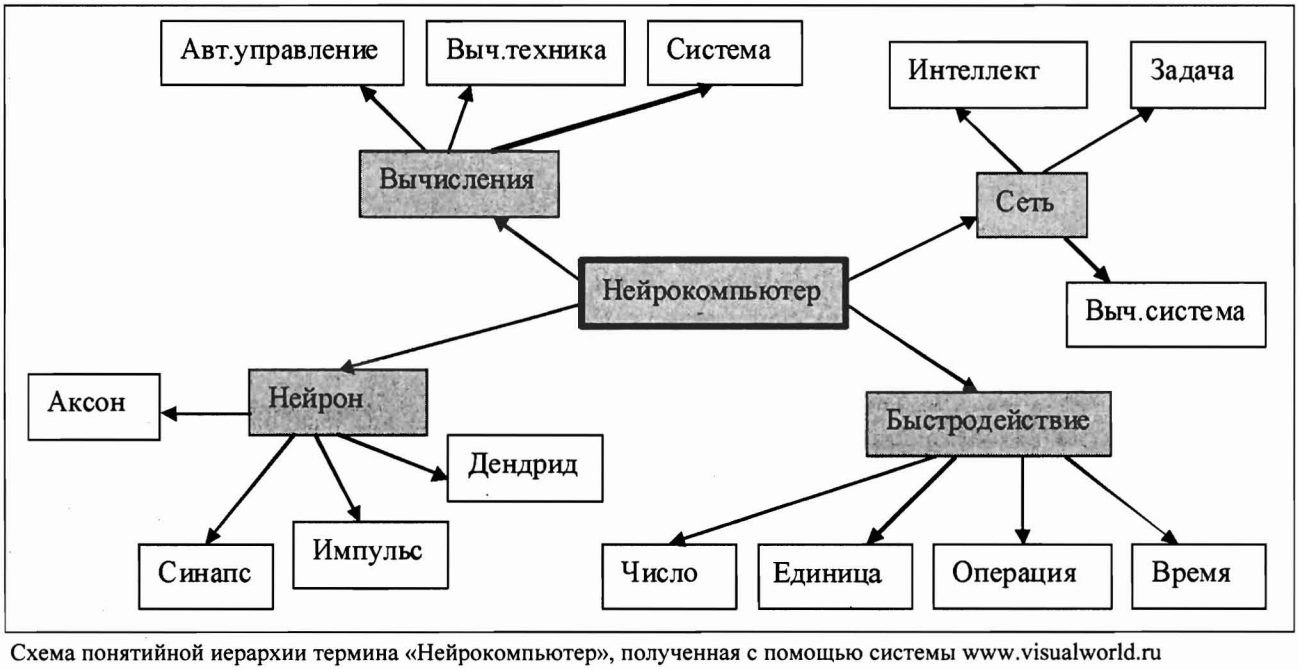

Схема понятийной иерархии термина «Нейрокомпьютер», полученная с помощью системы www.visualworld.ru

Таким образом, если значения терминов понятийного окружения выбранного понятия нам известны, например, из других предметных областей, то разобраться в новой предметной области (хотя бы получить общее представление о ней), становится возможных уже на основе визуального представления взаимосвязей понятий (рисунок).

На следующем этапе для каждого полученного термина, с помощью поискового сайта <www.visualworld.ru> отыскивались документы, наиболее полно раскрывающие их содержание. И на основе найденных документов составлялись соответствующие главы реферата.

В результате был получен реферат на тему «Нейрокомпьютер» общим объемом 37 страниц. В реферате отражены основные положения, касающиеся нейрокомпьютеров, нейронных сетей, их типов, методов обучения и пр.

Проведенные испытания предложенного метода показали его эффективность в качестве средства поддержки самоорганизующегося итерационного процесса терминологического изучения предметной области, который может быть использован для решения данного класса задач.

ЛИТЕРАТУРА

1. *Александров В.В., Андреева Н.А., Кулешов С.В.* Тенденции развития информационных систем: базы данных, базы знаний, онтологические, логистические, семиологические. Современные проблемы социально-экономического развития информационных технологий, сборник трудов (по итогам Международной научно-технической конференции), Баку, 2004.
2. *Александров В.В., Андреева Н.А., Кулешов С.В.* Тенденции развития информационных систем, IX Санкт-Петербургская Международная конференция «Региональная информатика – 2004», СПб, 22-24 июня 2004 г., материалы конференции, СПб, 2004.
3. *Александров В.В., Андреева Н.А., Кулешов С.В.* Визуальный динамический глоссарий – VISGLOSS, Материалы X Международной конференции и Российской научной школы «Системные проблемы надежности, качества информационных и электронных технологий (Инноватика-2005)», ч. 6. – М.: Радио и связь, 2005.
4. *Ножов И. М.* Процессор автоматизированного морфологического анализа без словаря. Деревья и корреляция // Диалог'2000. Труды конференции – Протвино, 2000, т.2, с. 284-290.

Поступила 15 января 2008 г.

System that Makes Conceptual Hierarchy for Associative Text Search

N.A. Andreeva, P.P. Kokorin

This paper proposes the method for building of conceptual hierarchy for associative text search. Description of software system for e-Leanring purposes based on this proposed method is shown. Infological search system <//www.visualworld.ru//> implementing associative text searching algorithm was used as the basis platform of the developed system. The system supports iterative self-organizing Internet based learning process.
Test results show real efficiency of proposed approach allowing solving the text search problem.