УДК 681.32

О.И. САФОНОВ

КОМПЬЮТЕР КАК НОВЫЙ КАНАЛ СВЯЗИ И ПРОБЛЕМА ПОИСКА ИНФОРМАЦИИ

Компьютерная революция, произошедшая в конце XX века, – это качественно новый этап цивилизации, в которой на первое место выходит компьютер как новый канал связи.

Интернет – это глобальная сеть компьютеров, объединяющая многие локальные, региональные и корпоративные сети и включающая сотни миллионов компьютеров, общающихся друг с другом на одном языке. Никто не управляет и не владеет системой целиком, но она связана таким образом, что позволяет работать с ней как с одной большой сетью. Сеть Интернет как феномен мирового масштаба в основном англоязычна, хотя она принимает и другие языки.

Компьютер, благодаря самому понятию информационных сетей, становится важнейшей частью глобального полилога, охватывающего в перспективе все человечество. В этой связи можно констатировать качественно иной тип культуры, выстроившийся вокруг персонального компьютера (ПК), который создал новый тип общения, основанный на свободном выходе личности в информационный космос.

Развитие глобальных компьютерных сетей началось в 80-е годы. В 1981 году в сети Интернет насчитывалось лишь 213 компьютеров, к концу 80-х годов количество подключенных к сети компьютеров возросло до 150 тысяч, однако наиболее быстрый экспоненциальный рост их количества произошел в 90-е годы. По данным Н.Д.Угриновича к началу 1999 года количество серверов в сети Интернет достигло 43 миллиона (Угринович, 1999).

Сегодня к Интернету имеет доступ в США около 120 млн. человек. В России Интернетом пользуются приблизительно 6 млн. человек. В Республике Беларусь – около полумиллиона, а в Витебске лишь – несколько тысяч человек. Здесь на некоторых почтовых отделениях открыли новую услугу: электронная почта по Интернету (e-meil), и это только начало.

Все большее количество учреждений образования получает доступ к Интернету. Ресурсы в сети создаются как крупными научными учреждениями и кампаниями, так и отдельными пользователями. В настоящее время на сотнях миллионов компьютеров, подключенных к Интернету, хранится громадный объем информации, накопленный человечеством на всю его историю; в связи с этим возникла острая потребность в обучении работе в Интернете.

Актуальность нашей работы обусловлена необходимостью научить ученика и студента не потеряться в мире информации, ведь в последнее время резко возросла роль поиска информации с использованием персонального компьютера (ПК). Это связано со следующими причинами:

- широким внедрением ПК во все (или почти все) сферы человеческой деятельности. ПК перестал быть инструментом для «избранных», а стал инструментом и хранилищем информации для многих. Причем хранилищем огромным. Сейчас не редкость встретить у домашнего пользователя винчестер в несколько гигабайт.
- ростом потока информации, которая вводится в ПК. Сканеры, качественные программы распознавания графических образов, а в последнее время и программы распознавания речи вносят свою немалую лепту.
- популярностью CD и DVD. Издано множество всевозможных энциклопедий, сборников текстов различной тематики (юридические справочники, рефераты, электронные издания журналов и газет).
- Интернет становится не просто «всемирной паутиной», но мощным экономическим рычагом влияния на состояние дел в мире. Постепенно экономика развитых стран становится Интернет-зависимой.
- Многие торговые сделки уже сейчас совершаются в Интернете. Таким стал I туризм. На ряде американских сайтов можно заказать авиабилеты в гостиницах ряда стран.
- Выход в ресурсы Интернет часто используется для работы с каталогами различных библиотек: крупнейшей в мире Библиотеки Конгресса США (http://www.lok.gov) или библиотек западных университетов, или Национальной Российской библиотеки.

- Получение свежей периодики. Полные тексты газетных и журнальных статей, scripts радио- и телепередач и т.д. Например, из бесплатных это 1) «USA Today»: http://www.usatoday.com; 2) «College Newspaper»:
- http://www.bowdoin.com; 3) новости от CNN: http://www.cnn.com; 4) прогноз погоды: http://www.weather.com и др.
- Информация из официальных источников. Можно, например, узнать, какие законы регулируют систему образования в любой стране, о чем говорилось на последних сессиях парламента и т.д.
- Информация из неофициальных источников. Например, можно узнать, как выглядят планы уроков учителей за рубежом. Каждый человек по договоренности со своим провайдером услуг Интернет может размещать в Сети страницы с различной информацией.
- Уже стало достаточно широким, а в скором времени, будем надеяться глобальным использованием сети Internet. О многом говорит только тот факт, что к 2015 году предполагают, что все ныне существующие книги будут переведены в электронный вид, а вновь издаваемые будут сразу появляться в обычном (бумажном) и электронном представлении. Уже сейчас многие журналы и газеты издаются в бумажном и электронном виде. Если еще сюда добавить переписку по E-mail, не потерявшие актуальности телеконференции. Gopher, ftp-архивы документов, то становится понятным насколько безбрежен океан документов. Надо только уметь им воспользоваться.

Уже сегодня стало понятно, что возможности Интернета безграничны: использование Интернет становится глобальным. Научить поиску в Интернете – наша задача.

К сожалению, ситуация сейчас такова, что информационному поиску практически не учат ни в школе, ни в вузах. В вузе методику преподавания информатики начали вести лишь в 1987/1988 учебном году, т.е. преподавание курса шло параллельно с его разработкой. Информатика – межпредметная дисциплина, это как бы интердеятельность (А.И.Бочкин, 1998, с.9).

Требования современности таковы, что необходимы новые средства организации доступа к информации, решение этих задач можно отнести к разряду систем искусственного интеллекта.

Основной задачей, возникающей при работе с полнотекстовыми базами данных, является поиск документов по их содержанию. Ставшие традиционными средства контекстного поиска по входящим в документ словам, представленные, в частности, поисковыми машинами в Internet, зачастую не обеспечивают адекватный выбор информации по запросу пользователя.

Главная проблема, по нашему мнению, заключается в сложности точной формулировки запроса. Это может быть связано с рядом причин:

недостаточным знанием терминологии предметной области, наличием в языке многозначных, синонимичных и омонимичных слов или даже слов с орфографическими ошибками в написании искомых слов, которые могут встречаться как в текстах, так и в самом запросе.

Более того, иногда пользователь не знает точно, какую именно информацию ему хотелось бы получить, поскольку имеет лишь общее представление о предмете поиска. Так, например, пытаясь расширить свои познания в области компьютерной лингвистики, на поисковом сервере Alta Vista вы получите список из сотен тысяч документов, содержащих слова «computer» и «linguistic». А хотелось бы получить расклассифицированный по тематическим группам материал,

отражающий, к примеру, хронологию событий в разработке этой области знаний. А часто необходим именно тот материал, о присутствии которого пользователь вообще не догадывается.

И хотя на сегодняшний день уже существует немало средств, способных помочь в извлечении информации, абсолютно адекватного еще не найдено.

Так, например, лидер мирового рынка СУБД Oracle уже снабдил разработчиков информационных систем рядом передовых технологий. Речь идет о картридже interMedia Text, входящем в состав СУБД ORACLE81, при использовании которого обработка текста сочетается со всеми возможностями, предоставленными пользователю Oracle для работы с реляционными базами данных. В частности, при написании приложений стало возможно использовать SQL с развитым языком запросов к полнотекстовой информации.

Не останавливаясь подробно на всех уникальных возможностях interMedia Text, информацию о которых можно получить на сайтах www. oracle. com и www.oracle.ru, отметим, что, к сожалению, большинство из них оказывается доступными в полной мере лишь хорошо владеющим английским языком. Адаптацией технологий Oracle к русскоязычным полнотекстовым базам данных занимаются специалисты компании «Гарант-Парк-Интернет». Продукт этой компании под названием Russian Context Optimizer (RCO) предназначен для совместного использования с картриджем interMedia Text. Рассмотрим подробнее эту интересную систему поиска.

Перенос технологий обработки текста с языка на язык не сводится к простой замене лингвистического наполнения. Поэтому реализация совместимости interMedia Text и RCO потребовала включения ряда оригинальных алгоритмов, отличных от используемых Oracle. Необходимо признать, что отсутствие на рынке некоторых важных компонентов лингвистической поддержки, в частности семантического словаря русского языка (который уже существует для английского), также заставляет разработчиков искать новые решения.

В основе технологий Oracle лежит использование семантического словаря английского языка – тезауруса, содержащего около полумиллиона слов, раскласcифицированных по тематическим категориям и синонимическим рядам: для каждого слова установлены его синонимы, более общие и более частные понятия, а также «родственные» слова, часто имеющие с ним смысловую связь в тексте.

Например, слову «лингвистика» соответствуют синонимы «языкознание» и «языковедение», а сам термин «лингвистика» относится к тематической группе, представленной более общим понятием «филология». В то же время «лингвистика» выступает в качестве тематической категории для ряда более частных понятий – «фонетика», «лексика», «морфология», а также имеет родственные по смыслу словосочетания: «наука о языке», «языковедческая дисциплина». В данном случае предложен термин, который имеет синонимы, вообще же синонимия – редкое и вредное в терминологии явление, поэтому, чтобы воспользоваться данной иерархической системой нужно задавать обычные (но высокочастотные в данной области, ключевые) слова, а не термины.

В целом иерархия категорий, представленных в тезаурусе, насчитывает до семи уровней вложенности и включает несколько тысяч тем по основным отраслям знаний. Абсолютно очевидно, что использование тезауруса в interMedia Text может оказать неоценимую помощь при контекстном поиске за счет расширения слов запроса различными видами близких по смыслу слов.

Наиболее примечательной оказывается способность interMedia Text проводить тематический анализ текста на английском языке. Текст каждого документа

подвергается процедурам лингвистической и статистической обработки, в результате чего определяются его ключевые темы.

Относя каждое слово текста к соответствующим разделам тезауруса и учитывая частоту встречаемости слов, interMedia Text может выделить до 16 главных тем документа.

Классификация документов по темам может оказать большую помощь при поиске, например, в случае, если пользователь затрудняется точно подобрать ключевые слова или же если он хочет сузить область поиска, уточнив тематику, по которой следует искать документы. Поиск по теме обладает более высокой точностью и полнотой по сравнению с простым контекстным поиском. Так, если последний находит все документы, содержащие заданные слова, то тематический поиск показывает лишь те документы, в которых словам запроса соответствует одна из ключевых тем. Кроме того, он позволяет найти документы, вовсе не содержащие слов из названия заданной темы, однако имеющие к ней отношение: по запросу «lower life forms» могут быть найдены документы, содержащие слова «bacteria» и «viruses».

В отличие от interMedia Text, RCO 3.0 пока не обладает мощным тезаурусом с иерархией категорий, который бы делал возможным глубинное обобщение информации в ходе тематического анализа. И если в тексте встречаются слова «компьютерный» или «компьютеризация», то RCO не сможет отнести документ к теме «вычислительная техника». Тем не менее RCO позволяет анализировать содержание текста, ставя в соответствие документу список его ключевых тем. Только RCO не выбирает названия тем из тезауруса, а выявляет их в тексте.

Используемое в RCO лингвистическое обеспечение позволяет приводить к нормальному виду все грамматические формы слов русского языка, сводить воедино различные части речи, а также отождествлять близкие по смыслу словосочетания. К примеру, выражения типа «языковые игры в русистике» и «несколько языковых игр» будут рассмотрены как одна и та же смысловая сущность.

Дополнительно RCO исключает из числа тем общеупотребительные слова, не несущие самостоятельной смысловой нагрузки или обладающие слишком широким значением. Так, слова «структура» и «стихотворный» сами по себе не могут характеризовать тему текста, но они могут входить в название темы, выраженной сочетанием с другим словом: «Структура стихотворного текста».

Другой замечательной способностью, которой обладает как RCO, так и interMedia Text, является автоматическое реферирование текста, которое происходит в ходе тематического анализа. При этом по каждой из выделенных тем выстраивается тематическое резюме, а также реферат текста.

Резюме формируется из фрагментов текста, причем если interMedia Text опирается на формальную разметку (наподобие html-тегов), то алгоритмы RCO способны самостоятельно членить текст на группы предложений, связанных общностью содержания, – сверхфразовые единства. В тематические резюме включаются лишь наиболее представительные, информативные фрагменты по соответствующим темам, в то время как общий реферат строится из фрагментов по всем главным темам документа. Визуализация списка ключевых тем и резюме при просмотре найденных документов ускоряют выбор требуемой информации. Так, даже взгляд на небольшой реферат может подсказать, следует ли читать документ полностью.

Однако отсутствие тезауруса в RCO не позволяет пока задействовать все возможности расширения запроса при контекстном поиске документов анало-

гично тем, которые доступны для английского языка при работе с interMedia Text.

Например, невозможно расширение слов запроса синонимичными, более общими или более частными, родственными по смыслу понятиями. Однако взамен этого RCO 3.0 обладает уникальной способностью, отсутствующей в interMedia. В отличие от предопределенных и очевидных связей, которые обычно задаются в тезаурусе, RCO устанавливает смысловые связи между темами, выявляя их в тексте динамически, так что большинство из них оказываются уникальными для каждой коллекции документов.

Так, RCO позволяет найти совокупность тем, связанных со словами запроса по смыслу в базе данных. Эта возможность оказывается полезна прежде всего аналитику, ведущему мониторинг событий, связанных с интересующей темой. Она позволяет определить «смысловое окружение» темы в коллекции документов и, уточнив запрос, выбрать требуемую информацию. Например, в ответ на запрос «лингвистика» можно получить следующий список тем: «развитие лингвистики», «история лингвистических учений», «государственная программа изучения», «русская лингвистика».

Остановимся кратко еще на одной полезной возможности – функции нечеткого поиска, позволяющей расширить запрос словами, близкими по написанию. Нечеткий поиск целесообразно применять при поиске слов с опечатками, а также в тех случаях, когда возникают сомнения в правильном написании, – фамилии, названия организации и т. п.

Понятно, что нечеткий поиск в interMedia Text не работает с русскоязычным текстом. Поэтому в RCO 3.0 реализован оригинальный алгоритм, использующий систему быстрого ассоциативного доступа к списку слов, содержащихся в документах, — в итоге можно найти слова по любым цепочкам составляющих их букв. Кроме того, RCO позволяет обнаружить нужное выражение по лексикографически близким словам (из группы документов, по которым ведется поиск), отличающимся заменами, пропусками и вставками символов. Например, по запросу «инкомбан», «инкобанки», «виткомбанки» обнаружится искомый «Инкомбанк». При этом допустимая близость найденных слов к запросу может задаваться при поиске.

Как видно, возможности поиска RCO 3.0 оказываются достаточно мощными, но не идеальными. Возможно, в следующую версию RCO войдут полноценный тезаурус русского языка, если таковой будет создан лингвистами, и тогда отпадет необходимость в подобного нашей работах, но пока мы должны искать наиболее оптимальные варианты поиска информации.

Чтобы нивелировать выявленные нами и отмеченные ранее недостатки, мы разработали технологию, предназначенную для обучения поиску некоторого содержания в большом массиве текстовых документов.

Таким образом, именно новые компьютерные технологии добавляют к традиционному образу культуры новое измерение.

ЛИТЕРАТУРА

1. ***Угринович Н.Д.*** Основы Интернет // Информатика и образование. – 1999. – № 9.
2. ***Бочкин А.И.*** Методика преподавания информатики: Учеб. пособие. – Мн.: Выш. шк., 1998.
3. ***Быкадоров Ю.А., Кузнецов А.Т., Павловский А.И.*** Информатика: Учеб. пособ. – Мн.: Нар. асвета, 1995.

SUMMARY

The Internet as a new channel of communication is described in the article. It is proved that it is necessary to adapt the global net to a Russian- speaking user.