Проблема поиска знаний и инструменты ее решения

Ольга Валерьевна Лебедева
магистр инженерных наук
Рижский технический университет
ул. Межа 3/4-533, г. Рига, Латвия, LV-1048, (+371)29808607
o.lebedeva@inbox.lv

Аннотация

В данной статье рассматривается проблема поиска знаний и предлагается обзор существующих инструментов ее решения. В работе дана классификация систем поиска знаний и описан принцип их работы. Выделено три типа систем поиска знаний: поисковые машины знаний, вопросно-ответные системы, комбинированные системы. На основе результатов тестирования систем каждого типа (TrueKnowledge, START, Wolfram|Alpha, Freebase, Blurtit, Allexperts, Ответы@mail.ru, Answers) выявлены преимущества и недостатки их работы.

This paper defines the knowledge search problem and presents the overview of existing solutions. The classification of systems for knowledge search is given here and the working principle of these systems is described. Three types of systems for knowledge search are marked out: knowledge search engines, Q&A websites, combined systems. The advantages and disadvantages of each system type were revealed due to testing results of the following systems: TrueKnowledge, START, Wolfram|Alpha, Freebase, Blurtit, Allexperts, Ответы@mail.ru, Answers.

Ключевые слова

знания, поиск знаний, система поиска знаний, поисковая программа, вопросно-ответная система;

knowledge, search for knowledge, system for knowledge search, search engine, Q&A website.

Введение

По мере общественного развития в качестве источника прибыли все чаще выступают знания, инновации и способы их практического применения. Приобретение новых знаний, информации, умений, навыков, а также их постоянное обновление и развитие становятся главными приоритетами работников в условиях современной экономики [1]. Увеличение средней продолжительности жизни, и, как следствие, повышение возраста выхода на пенсию, вызывают необходимость не только постоянно повышать квалификацию, но и несколько раз в течение жизни менять профессию. Поэтому сегодня многие ученые занимаются вопросами качественной подготовки квалифицированных специалистов: разрабатываются инновационные методы обучения (модульное, проблемное, дистанционное, совместное, мобильное, и др. [2-4]); создаются специализированные обучающие системы (*Moodle* [5], *Blackboard* [6], и др.); разрабатываются курсы повышения квалификации и/или переквалификации (*continuing education, lifelong learning*, и др. [7]). Следовательно проблема нехватки и, соответственно, поиска знаний актуальна на протяжении всей сознательной жизни человека.

В настоящее время проблемой поиска знаний занимаются специалисты в области управления личными и корпоративными знаниями (Knowledge Management), т.к. это позволит сократить временны*е* и трудозатраты на выполнение задач и принятие решений, как на рабочем месте, так и в повседневной жизни [8, 9]. Особенно остро стоит проблема поиска знаний в Интернете, т.к. при постоянно растущем количестве информации в сети, извлечь действительно нужную и важную информацию становится все труднее [10]. Поэтому поиском решения данной проблемы также занимаются разработчики поисковых систем, таких как Google, Яндекс, Рамблер [11-14].

Кроме того, проблема поиска знаний актуальна и в сфере образования. Современные образовательные стандарты помимо всего прочего включают в себя перечень ключевых образовательных компетенций, в том числе так называемые информационные компетенции [15]. Информационная компетентность это «*способность и умение самостоятельно искать, анализировать, отбирать, обрабатывать и передавать необходимую информацию при помощи устных и письменных коммуникативных информационных технологий*» [16]. Другими словами, квалифицированный специалист должен уметь самостоятельно работать с информацией, что на сегодняшний день практически невозможно без использования компьютерных технологий. Поэтому одна из задач современного образования состоит в том, чтобы научить человека эффективно работать с информацией. Например, в помощь студентам научной библиотекой Национального исследовательского ядерного университета «МИФИ» было подготовлено учебно-методическое пособие «Основы информационно-библиографической культуры», с целью сформировать коммуникативные навыки, в том числе освоить способы самостоятельного поиска информации; освоить использование информационных технологий в образовательной деятельности, помочь в овладении навыками информационно-поисковой работы для написания курсовых, дипломных и других научных работ [17].

Однако, не смотря на актуальность проблемы поиска знаний, она все еще остается малоизученной, а предлагаемые варианты ее решения далеки от совершенства. Поэтому цель данной статьи – описать проблему поиска знаний, классифицировать существующие инструменты ее решения (системы поиска знаний), проанализировать результаты их работы, а также рассмотреть возможность использования данных систем в учебном процессе.

Понятие поиска знаний

В данной статье под *знаниями* понимается осознанная *информация*, дополненная личным опытом, эмоциями, убеждениями, хранящаяся в памяти человека, и связанная с другими знаниями. Иными словами, знания нельзя прочитать в книге, но можно прочитать информацию и на ее основе сформировать знания у себя в голове. Поэтому поиск знаний в прямом смысле этого слова не возможен, за исключением случая, когда человек ищет знания в своей личной памяти, т.е. пытается что-то вспомнить. Следовательно, **поиск знаний представляет собой поиск конкретной и четко сформулированной информации, которой человеку достаточно для формирования знаний.**

Любой поиск включает в себя следующие этапы:

- определение искомой информации, т.е. что именно необходимо найти;
- определение источника информации, т.е. где имеет смысл искать;
- определение способа поиска, т.е. как искать, учитывая специфику искомой информации и место нахождения;
- непосредственно сам поиск.

При поиске информации для формирования знаний, человек старается использовать несколько источников, чтобы ознакомиться с различными взглядами на предмет поиска, а также иметь возможность удостовериться в истинности найденных фактов. Таким образом, параллельно запускаются несколько процессов поиска, результаты которых обобщаются с целью обнаружения и последующего устранения противоречий. В итоге на выходе данного процесса человек получает достоверную информацию, хотя стоит отметить, что истинность найденных фактов напрямую зависит от источников информации, которые использовались при поиске.

На сегодняшний день все более и более популярным источником информации становится сеть Интернет. Согласно статистике [18] количество пользователей Интернета неуклонно растет, причем как в развитых странах, так и в развивающихся. Поиск знаний в Интернете несколько отличается от общей схемы поиска информации представленной выше. Это связано с тем, что место нахождения информации изначально определено. Кроме того тип источника информации четко определяет способы поиска.

Можно выделить три основных этапа поиска знаний в Интернете: дебют, миттельшпиль и эндшпиль, проводя аналогию с шахматной партией [10].

Дебют предполагает определение специфики искомой информации и зоны поиска. Поскольку количество информации в Интернете огромно, найти что-либо конкретное без помощи поисковой системы не возможно, за исключением случая, когда пользователю известен точный адрес веб-страницы с необходимой информацией. Поэтому на первом этапе поиска задача пользователя состоит в том, чтобы выбрать круг поиска: искать на тематическом веб-сайте или в Интернете вообще; искать только файлы определенного типа или группы типов, например, текстовые; и т.д. На этапе миттельшпиля пользователь формулирует варианты запросов к поисковой системе. Запрос определяет критерии поиска, чем точнее он сформулирован, тем точнее будут результаты поиска. Запрос может иметь форму словосочетания или вопроса на естественном языке, кроме того некоторые поисковые системы предлагают возможность включить в запрос логические (AND, OR, NOT и др.) и/или контекстные операторы (для задания определенных параметров поиска, например, порядка расположения слов). Заключительный этап – эндшпиль – представляет собой обработку результатов поиска, а именно, отбор нужной информации.

Среди перечисленных этапов отсутствует процесс поиска как таковой. Данный факт объясняется тем, что Д.В. Ландэ [10] рассматривает поиск знаний в Интернете с точки зрения действий, производимых пользователем, тогда как сам поиск осуществляет поисковая система. Таким образом, процесс поиска знаний в Интернете состоит из четырех основных этапов:

1. Определение искомой информации и зоны поиска;
2. Формулировка вариантов запроса;
3. Непосредственно поиск;
4. Обработка результатов поиска.

На рисунке представлена схема процесса поиска знаний, где первый этап разбит на подпроцессы: «Определение искомой информации» и «Определение зоны поиска»; а последний этап – на подпроцессы: «Обобщение результатов поиска» и «Устранение противоречий».

Далее в статье описываются принципы работы существующих систем поиска знаний, а также приводится их классификация.

Системы поиска знаний

Как уже упоминалось выше, поиск знаний – это поиск определенной информации, а именно, конкретной и четко сформулированной. Значит система поиска знаний – это система, которая ищет такую информацию, или же система, которая ищет любую информацию о предмете, затем обрабатывает ее, и возвращает результат в виде конкретной и четко сформулированной информации. Эффективность поиска во многом зависит от формулировки запроса, т.к. именно он задает параметры поиска. Запрос на естественном языке наиболее удобен для человека. Например, поиск по ключевым словам имитирует поиск в энциклопедии, используя предметный указатель, а поиск, заданный с помощью вопросительного предложения, напоминает диалог между людьми. Поэтому системы поиска знаний предоставляют пользователю возможность вводить запрос как в виде словосочетания на естественном языке, так и в виде вопроса.

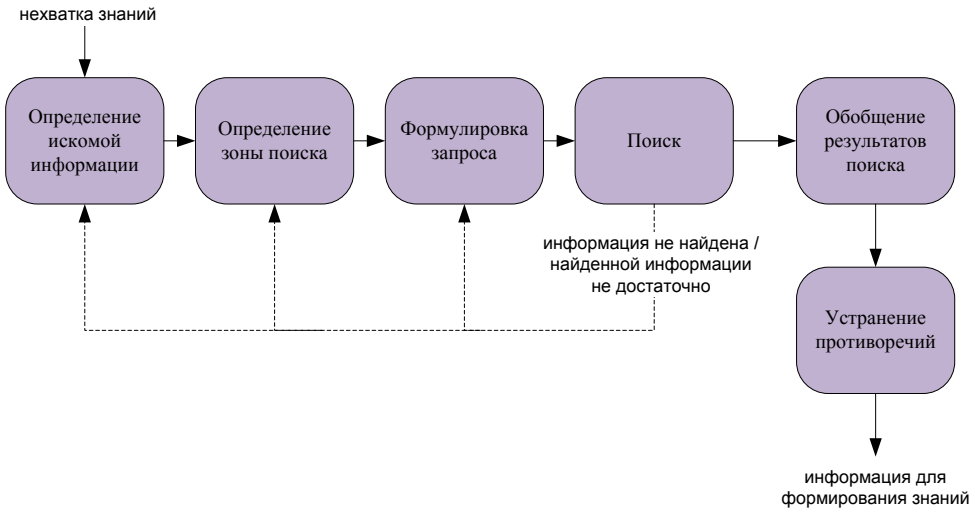

Рис. Схема процесса поиска знаний в Интернете

Системы поиска знаний можно разделить на три типа: поисковые машины знаний (knowledge search engines), вопросно-ответные системы (Q&A websites) и комбинированные системы. Данная классификация основана на различиях в принципах работы систем поиска знаний. Поисковые машины знаний самостоятельно производят обработку запроса пользователя, поиск и генерацию ответа. В свою очередь, вопросно-ответные системы осуществляют данные действия посредством пользователя, т.е. одни пользователи публикуют на веб-сайте вопросно-ответной системы свои вопросы, а другие пользователи на них отвечают. Комбинированные системы объединяют в себе принципы работы предыдущих двух типов, т.е. работают и как поисковая машина знаний, и как вопросно-ответная система.

Поисковые машины знаний имеют собственную встроенную базу фактов, а также механизмы вывода, которые позволяют им генерировать ответ, основываясь на найденных фактах. Получив запрос пользователя, поисковая машина знаний формализует его, т.е. трансформирует в вид, понятный программе поиска. Затем осуществляется поиск в базе фактов. Если необходимые факты найдены, то поисковая машина знаний формирует ответ в виде текста на естественном языке (допустимы также графические элементы), основываясь на фактах и используя механизмы вывода. В результате пользователь получает конкретный ответ на свой

запрос, например, запрос: «*Какова температура плавления золота?*», ответ: «*1063°C*», причем пользователь может посмотреть на основании каких фактов был выведен данный ответ. Если в базе фактов необходимые факты не найдены, то поисковая машина знаний сообщает пользователю о том, что не может ответить на его запрос, или обращается к внешним источникам информации, например, ищет информацию в Интернете как обыкновенная поисковая система.

Вопросно-ответная система представляет собой веб-сайт, на котором как зарегистрированные, так и незарегистрированные пользователи могут задавать друг другу вопросы. Прототипом такой системы является Интернет-форум. Принцип работы следующий: пользователь адресует вопрос сообществу пользователей, его вопрос публикуется на веб-сайте вопросно-ответной системы, после чего каждый, кто может и хочет ответить, публикует свой ответ. Таким образом, пользователь получает несколько различных ответов на свой вопрос. Но есть веб-сайты (например, www.Allexperts.com), где на вопросы пользователей отвечают только эксперты. Здесь пользователь адресует свой вопрос конкретному эксперту, и может разрешить или запретить публикацию вопроса и ответа. Стоит отметить, что ответы вопросно-ответной системы в большинстве случаев представляют собой скорее мнения людей, нежели «голые» факты, лишенные субъективной оценки. Все вопросы и ответы сохраняются, поэтому часто пользователю достаточно просто воспользоваться функцией поиска, чтобы найти нужную информацию среди отвеченных вопросов.

Комбинированная система объединяет возможности поисковой машины знаний и вопросно-ответной системы. У нее есть как собственная встроенная база фактов, так и сообщество пользователей, готовых отвечать на вопросы. При обработке запроса пользователя комбинированная система сначала обращается к собственной базе фактов. Если запрос пользователя невозможно формализовать, т.е. система не понимает что нужно искать, или в базе фактов нет необходимой информации, то запрос переадресовывается сообществу пользователей.

В таблице 1 представлена классификация систем поиска знаний, даны их основные характеристики, а также приведены примеры систем, которые свободно доступны всем пользователям Интернета.

Таблица 1. Классификация систем поиска знаний

Тип	**Основные характеристики**					**Примеры**
	Встроенная база фактов	Сообщество пользователей	Форма запроса	Форма ответа	Время ожидания	
Поисковые машины знаний	+	–	В, С	ОФ	min	TrueKnowledge [19] START [20] Wolfram|Alpha [21] Freebase [22]
Вопросно-ответные системы	–	+	В, С[1]	СФ	неизвестно и неограниченно	Blurtit [23] Allexperts [24] Ответы@mail.ru [25]
Комбинированные системы	+	+	В, С	ОФ[2], СФ[3]	[min, ∞]	Answers [26]

[1] при поиске среди отвеченных вопросов

[2] в случае, когда комбинированная система работает как поисковая машина знаний

[3] в случае, когда комбинированная система работает как вопросно-ответная система

Пояснения к таблице 1:

- В – вопросительное предложение на естественном языке;
- С – словосочетание на естественном языке (ключевые слова);
- ОФ – строго определенная форма ответа;
- СФ – свободная форма ответа;
- min – время ожидания ответа минимально (в основном зависит от скорости обработки запроса).

Анализ работы систем поиска знаний

В ходе исследования систем поиска знаний, перечисленных в таблице 1, были выявлены преимущества и недостатки характерные для каждого типа. Вследствии изучения принципа работы поисковых машин знаний, удалось установить, что пользователь всегда получает четко сформулированный ответ на свой запрос, однако, недостаток фактов во встроенной базе и несовершенство механизма вывода часто не позволяет поисковой машине знаний понять запрос пользователя и правильно на него ответить. В свою очередь принцип работы вопросно-ответных систем практически исключает возможность контроля качества ответов, поэтому ответственность за истинность, корректность и полноту ответа несет лишь пользователь, отвечающий на вопрос. Помимо изучения принципа работы систем поиска знаний, был разработан и проведен тест с целью определить результативность работы данных систем. Тест состоит из семи вопросов:

1. Что такое северное сияние?
2. Сколько государств в мире?
3. Когда закончилась 100-летняя война?
4. Как приготовить яблочный пирог (штрудель)?
5. Как работает радиоприемник?
6. Почему листья растений зеленые?
7. Из-за чего началась Первая мировая война?

Задача систем поиска знаний состояла в том, чтобы правильно ответить на данные вопросы. Правильным считался ответ, суть которого совпадала с эталонным ответом. Например, эталонный ответ на вопрос «*Сколько государств в мире?*» звучит так: «*На 23 марта 2012 года в Организации Объединенных Наций числится 192 страны, хотя количество независимых государств составляет 196*» [27]. Поэтому если в ответе системы поиска знаний фигурировало число 192 или 196, ответ считался правильным.

В таблице 2 даны результаты проведенного теста, используя поисковые машины знаний (ПМЗ), вопросно-ответные системы (ВОС) и комбинированные системы (КС). Ответы систем оценивались по следующей шкале: 0 – ответа нет; 1 – неправильный ответ; 2 – правильный ответ. Также определено количество правильных ответов К и результативность работы Р каждой системы в процентах.

Таблица 2. Результаты тестирования работы систем поиска знаний

Тип	Система	Вопросы							К	Р, %
		1	2	3	4	5	6	7		
ПМЗ	TrueKnowledge	2	2	2	0	0	2	0	4	**57**
	START	2	2	1	0	0	0	0	2	**29**
	Wolfram|Alpha	1	1	1	1	1	1	1	0	**0**
	Freebase	0	0	0	0	0	0	0	0	**0**

ВОС	Blurtit	1	2	0	0	2	0	2	3	**43**
	Allexperts	2	1	0	2	0	2	0	3	**43**
	Ответы@mail.ru	2	1	1	1	2	2	2	4	**57**
КС	Answers	2	1	0	0	0	2	0	2	**29**

Результативность работы системы рассчитывалась по формуле:

$$P = \frac{K}{\textit{общее число вопросов}} \times 100\%$$

Анализ ответов систем также помог выявить их некоторые преимущества и недостатки. Например, вопросно-ответные системы почти всегда предлагают пользователю несколько вариантов ответа, что дает возможность ознакомиться с различными мнениями и сформировать свое собственное, однако формулировка ответа бывает неоднозначной и/или некорректной, что может привести к формированию неверных знаний.

Ниже перечислены все выявленные в ходе исследования преимущества и недостатки систем поиска знаний.

Преимущества:

- поисковые машины знаний моментально возвращают ответ пользователю, как правило основываются на надежных источниках информации, и генерируют ответ в строго определенной структурированной форме;
- вопросно-ответные системы умеют отвечать на так называемые сложные вопросы (вопросы, на которые нет одного единственного правильного ответа), предлагают разные варианты ответов (что дает возможность пользователю ознакомиться с различными мнениями и сформировать свое собственное), а также обеспечивают возможность диалога между задающим вопрос и отвечающим на него;
- комбинированная система обладает всеми преимуществами поисковых машин знаний и вопросно-ответных систем.

Недостатки:

- поисковые машины знаний часто не понимают или неправильно понимают запрос пользователя, а также им не хватает знаний (фактов и механизмов вывода) для того, чтобы сформулировать ответ;
- вопросно-ответные системы часто дают основания сомневаться в истинности ответа, точность и корректность формулировки ответа зависит только от отвечающего (требований к форме ответа нет), кроме того время ожидания ответа не ограничено, т.к. неизвестно когда появится пользователь, который может и хочет ответить на данный вопрос;
- комбинированная система не обладает недостатками поисковых машин знаний, т.к. запрос, который она не смогла понять или на который не смогла ответить, перенаправляется сообществу пользователей и кто-нибудь из них скорее всего на него ответит, однако при этом сохраняются все недостатки вопросно-ответных систем.

Как упоминалось выше, одна из задач современного образования – научить человека эффективно работать с информацией, в том числе быстро ее находить. Причем речь идет не о любой информации, а об информации, на основе которой можно сформировать знания. Для поиска такой информации рациональней применять именно системы поиска знаний, а не информационно-поисковые системы, т.к. последние возвращают пользователю не конкретную информацию о предмете поиска, а список ссылок на Веб-страницы, на которых потенциально содержится информация о предмете. Так, при поиске энциклопедических знаний рекомендуется

применять поисковые машины знаний, т.к. они предоставляют пользователю достоверную и четко сформулированную информацию. При поиске практических знаний, таких как результаты применения методологии, *best practices*, и т.п., рекомендуется применять вопросно-ответные системы, т.к. они дают пользователю прямой выход на экспертов в нужной области. Очевидно, что наиболее универсальным вариантом является комбинированная система, т.к. она может работать как поисковая машина знаний и как вопросно-ответная система. Однако, стоит отметить, что существующие системы поиска знаний имеют ряд существенных недостатков, что ставит под сомнение возможность их применения в учебном процессе: если пользователь получит ложную или нечетко/некорректно сформулированную информацию, то в его сознании скорее всего сформируются ложные знания.

Заключение

В условиях постоянно растущего количества информации все большую актуальность приобретает проблема извлечения из информационного потока действительно нужной и полезной информации, другими словами, проблема поиска знаний. Актуальность данной проблемы вызвала интерес к поиску ее решения со стороны разработчиков поисковых систем и технологий обмена знаниями, поэтому уже сегодня существуют различные инструменты поиска знаний: поисковые машины знаний, вопросно-ответные системы и комбинированные системы. Рассмотренные в статье системы обладают не только преимуществами, но и существенными недостатками, а именно, часто не понимают запрос пользователя, возвращают неправильный или неподкрепленный фактами ответ. Наличие этих недостатков в значительной степени ограничивает возможности применения систем поиска знаний. Ведь система, которая дает конкретный, четко сформулированный ответ на запрос в виде словосочетания или вопросительного предложения на естественном языке, могла бы широко применяться в учебном процессе и на практике. Поэтому дальнейшие исследования в этом направлении будут посвящены разработке системы поиска знаний, которая бы не обладала недостатками существующих систем, а также была адаптирована к применению в учебном процессе.

Литература

1. Центр новых информационных технологий при аэрокосмическом университете: Методология и технология электронного обучения [Электронный ресурс]. – URL: http://cnit.ssau.ru/do/index.htm (дата обращения 15.05.2012).
2. Голицына И.Н., Половникова Н. Л. Мобильное обучение как новая технология в образовании // Образовательные технологии и общество: - 2011. – Т. 14. – № 1. – С.241-252.
3. Морозов М.Н., Герасимов А. В., Курдюмова М.Н. Системы совместной учебной деятельности на основе компьютерных сетей // Образовательные технологии и общество: - 2009. – Т. 12. – № 1. – С.310-323.
4. Устюгова В.Н., Валитов Р. А. О процессе создания системы дистанционного обучения в Татарском государственном гуманитарно-педагогическом университете (ТГГПУ) // Образовательные технологии и общество: - 2010. – Т. 13. – № 2. – С.225-239.
5. Moodle company Web site [Электронный ресурс]. – URL: http://moodle.com (дата обращения 10.05.2012).

6. Blackboard company Web site [Электронный ресурс]. – URL: http://www.blackboard.com (дата обращения 18.05.2012).
7. The Standard for Lifelong Learning IACET [e-resource]. – URL: http://www.iacet.org (дата обращения 17.05.2012).
8. Dorsey, P. Personal knowledge management [e-resource]. – URL: http://www.360doc.com/content/05/1228/22/2563_51065.shtml (дата обращения 16.02.2012).
9. Martin, J. Personal Knowledge Management. The Basis of Corporate and Institutional Knowledge Management// Managing Knowledge: Case Studies in Innovation. – Alberta: University of Alberta, faculty of Extension, 2000. – Vol.6.
10. Ландэ, Д.В. Поиск знаний в INTERNET. – Москва: Диалектика, 2005. – 12-14, 52-54 стр.
11. Как найти информацию: Как правильно искать в Google (1) [Электронный ресурс]. – URL: http://www.vsepoisk.ru/2009/02/google-1.html (дата обращения 05.05.2012).
12. Официальный сайт компании Яндекс: Технологии: Поисковые технологии [Электронный ресурс]. – URL: http://company.yandex.ru/technologies/query/index.xml (дата обращения 10.05.2012).
13. Энциклопедия поисковых систем: Принципы работы поисковой машины Рамблер [Электронный ресурс]. – URL: http://www.searchengines.ru/articles/004575.html (дата обращения 08.05.2012).
14. Google Company Web site: What we believe: Ten things we know to be true [e-resource]. – URL: http://www.google.com/about/company/philosophy/ (дата обращения 03.05.2012).
15. Хуторской А.В. Доклад "Определение общепредметного содержания и ключевых компетенций как характеристика нового подхода к конструированию образовательных стандартов" [Электронный ресурс]. – URL: http://www.eidos.ru/journal/2002/0423.htm (дата обращения 10.05.2012).
16. Федеральный государственный образовательный стандарт: Глоссарий: Компетентность информационная [Электронный ресурс]. – URL: http://standart.edu.ru/catalog.aspx?CatalogId=791 (дата обращения 11.05.2012).
17. Учебно-методическое пособие «Основы информационно-библиографической культуры». Центр информационно-библиотечного обеспечения учебно-научной деятельности [Электронный ресурс]. – URL: http://library.mephi.ru/icb2/book.html (дата обращения 15.05.2012).
18. International Telecommunication Union: Development: Market Information and Statistics [e-resource]. – URL: http://www.itu.int/ITU-D/ict/statistics/ (дата обращения 20.01.2012).
19. True Knowledge Web page: About us. – URL: http://www.trueknowledge.com/about (дата обращения 05.04.2012).
20. START Natural Language Question Answering System. – URL: http://start.csail.mit.edu (дата обращения 12.03.2012).
21. Wolfram|Alpha Web page: About. – URL: http://www.wolframalpha.com/about.html (дата обращения 18.04.2012).
22. Freebase Web page: About us. – URL: http://www.freebase.com/view/m/021ympy (дата обращения 15.02.2012).
23. Blurtit Web page: What is Blurtit. – URL: http://www.blurtit.com/about/what_is_blurtit (дата обращения 07.03.2012).
24. Allexperts Web page: About us. – URL: http://www.allexperts.com/central/service.htm (дата обращения 18.04.2012).
25. Веб-сайт Ответы@mail.ru. – URL: http://otvet.mail.ru/ (дата обращения 20.03.2012).

26. Answers.com Web page: About. – URL: http://wiki.answers.com/about/ (дата обращения 12.02.2012).
27. About.com: Geography: The Number of Countries in the World [e-resource]. – URL: http://geography.about.com/cs/countries/a/numbercountries.htm (дата обращения 15.04.2012).