

К ВОПРОСУ О РАЗРАБОТКЕ МЕТОДИКИ ПОИСКА НАУЧНОЙ ИНФОРМАЦИИ С ИСПОЛЬЗОВАНИЕМ СПЕЦИАЛИЗИРОВАННЫХ ПРОГРАММНЫХ КОМПЛЕКСОВ

Ахметгалеев Р.Р. Email: Akhmetgaleev630@scientifictext.ru

*Ахметгалеев Рустам Рамилевич – аспирант,
кафедра вычислительной техники, факультет информатики и вычислительной техники,
Ижевский государственный технический университет им. М.Т. Калашникова, г. Ижевск*

Аннотация: в статье рассматриваются проблемы поиска научной информации в условиях больших объемов данных сети Интернет, содержащих значительное количество шумовой информации. Приведены результаты анализа современных подходов и инструментов поиска научной информации, где изложены их основные недостатки. Особо подчеркивается взаимосвязь между семантическим пространством исследователя и эффективностью поиска научной информации. В контексте указанной взаимосвязи отмечается актуальность вопросов создания подходов по повышению эффективности поиска. В качестве одного из таких подходов предлагается методика поиска, основанная на использовании специализированной системы поиска. Кратко рассматриваются архитектура и принципы работы разрабатываемой системы.

Ключевые слова: системы поиска информации, методы индексации, НИР, релевантность поиска, научная работа.

TO THE QUESTION OF SCIENTIFIC INFORMATION SEARCHING METHOD DEVELOPMENT WITH USAGE OF SPECIALIZED SOFTWARE COMPLEXES

Akhmetgaleev R.R.

*Akhmetgaleev Rustam Ramilevich – Graduate Student,
DEPARTMENT OF COMPUTER ENGINEERING,
IZHEVSK STATE TECHNICAL UNIVERSITY M.T. KALASHNIKOV, IZHEVSK*

Abstract: the article deals with the scientific information searching problems in conditions of large volumes of data, which is containing a considerable amount of noise information and storing in the Internet. The results of the analysis of modern approaches and scientific information searching tools are presented, where their main disadvantages are stated. Particular emphasis is placed on the relationship between the researcher semantic space and the effectiveness of the scientific information search. In the context of this relationship, the relevance of issues of creating approaches for improvement searching effectiveness is noted. As one such approach, a search technique based on the usage of a specialized search system is proposed. Briefly discusses the architecture and working principles of the specialized search system which is under development now.

Keywords: information retrieval systems, indexing methods, research, search relevance, scientific work.

УДК 004.91

В настоящее время быстрое развитие информационных технологий способствует стремительному увеличению объема информации, хранящейся в сети Интернет. Так, например, результаты исследований компании International Data Corporation, ведущей свою деятельность в сфере анализа данных, показывают, что до 2020 года объем информации будет увеличиваться более чем в два раза каждые два года [1]. Кроме того, схожая тенденция лавинообразного роста информации наблюдается и в науке. В своих исследованиях аналитики Л. Борнманн и Р. Мутц делают выводы о том, что невозможно точно подсчитать объемы научной информации, однако можно оценить темп ее увеличения, который составляет около 8-9% в год [2]. Данный показатель эквивалентен удвоению мировых научных результатов примерно каждые девять лет. На графике 1, представленном ниже, отражены темпы роста объема научных трудов.

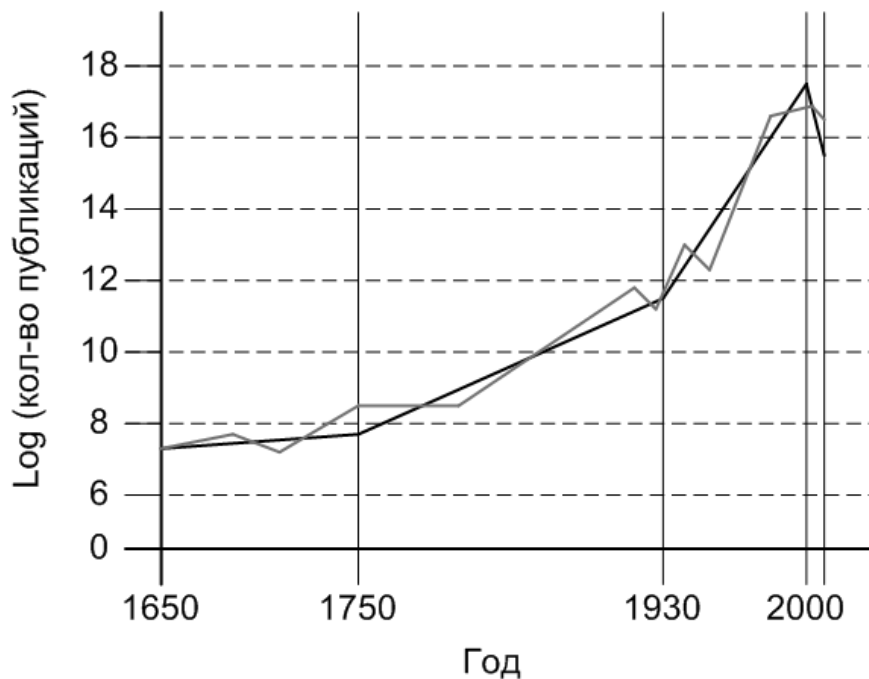


Рис. 1. Данные по оценке примерного объема научной информации

Значительно увеличивающиеся объемы данных в сети Интернет приводят к тому, что необходимая научная информация теряется в больших массивах неконструктивной, шумовой или дублирующейся информации. Еще одной причиной усложнения поиска научной информации является то, что современные поисковые машины оптимизированы для поиска фактологической и справочной информации. Стоит отметить, что в настоящее время эти задачи решаются на достаточно высоком уровне. Однако возможностей тех же самых поисковых машин недостаточно для качественного поиска научной информации. Это обусловлено, в первую очередь, тем, что исследователь стремится найти научные труды, способствующие достижению его целей, в условиях недостаточно широкого семантического пространства в определенной сфере знаний. Перечисленные факторы заметно снижают эффективность и качество информационного поиска.

Отличительной особенностью современных информационных поисковых систем (ИПС) является, то что пользователю предоставляется одно поле, куда он мог бы ввести текст своего запроса. В случае с поиском фактологической и справочной информации такой подход в дизайне поисковых систем давно себя зарекомендовал с положительной стороны. При этом точность результатов поиска, во многом зависит от того, насколько полно текст поискового запроса отражен в найденных документах. Однако в случае поиска научной информации более важным является семантическое соответствие найденного текста и запроса, поскольку ключевые слова запроса подбираются в соответствии с целями исследователя и ожидается, что в найденных текстах будет отражена информация необходимая для решения поставленных задач. Помимо этого, во многих поисковых машинах отсутствуют такие функции, как поиск по компонентам научной работы и прочим атрибутам. В настоящее время предпринимаются попытки адаптировать существующие ИПС под специфику поиска научной информации посредством добавления дополнительных критериев поиска. Однако они существенно не повышают эффективность поиска научной информации. В связи с этим развиваются специализированные ИПС, предназначенные для организации эффективного поиска научной информации. Тенденция развития специализированных ИПС подтверждается значительным количеством опубликованных трудов в данной области. Анализируя работы по специализированным ИПС была составлена обобщенная архитектура таких систем, она отображена на рисунке 2. Одним из главных компонент такой системы является модуль индексации. То, насколько точно будут выполнены задачи классификации документов, поиска по ключевым словам и терминам, а также задачи выделения терминов из текста, напрямую зависит от качества реализации модуля индексации и используемых в нем алгоритмов. Несмотря на развитие таких систем, проблемы построения эффективного процесса информационного поиска сохраняются, поскольку исследователь работает в ограниченном семантическом пространстве. При этом недостаточно проработаны механизмы, способные снизить барьеры между семантическим пространством исследователя и массивом информации, в которой происходит поиск. Это также является негативным фактором, снижающим эффективность поиска.

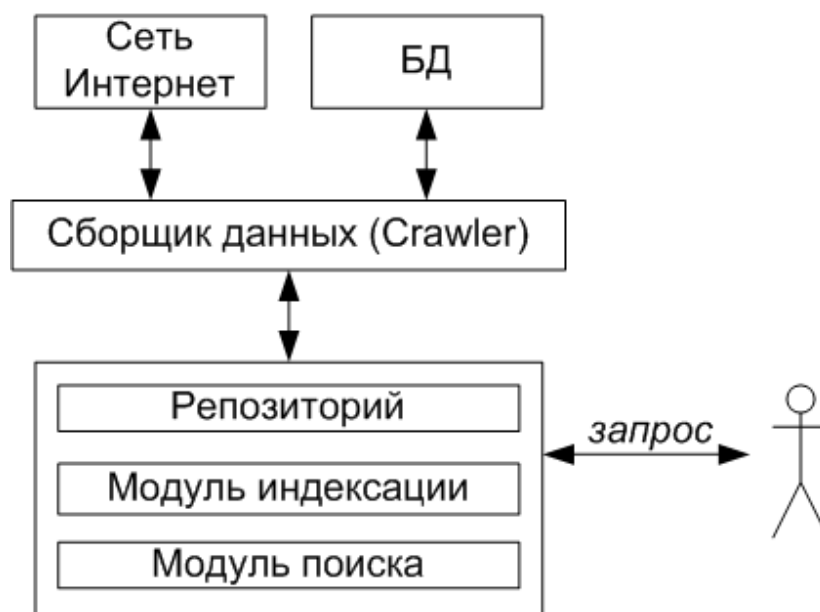


Рис. 2. Обобщенная архитектура систем поиска научной информации

На основе проведенного обзора и анализа специализированных ИПС и методов индексации текста были сделаны следующие выводы о недостатках существующих в современных специализированных системах поиска научной информации:

- используемые алгоритмы индексации в современных ИПС показывают не очень высокую степень точности в задачах извлечения терминов из текстов;
- не учитываются особенности поиска научной информации, которые заключаются в том, что поисковый запрос составляется в соответствии с семантическим пространством исследователя. При этом высока вероятность ситуации, когда в хранилище документов может присутствовать необходимая информация, однако она не попадает в результирующую выборку из-за того что семантическое пространство документа и запроса сильно различаются. Это в свою очередь приводит к необходимости тратить дополнительное время на поиск;
- отсутствуют механизмы поиска по компонентам научных работ.

Для устранения вышеизложенных недостатков предлагается методика проведения поиска научной информации с использованием специализированной ИПС, архитектура которой показана на рисунке 3. В рамках предлагаемой методики лежит итерационный подход по уточнению поисковых запросов.

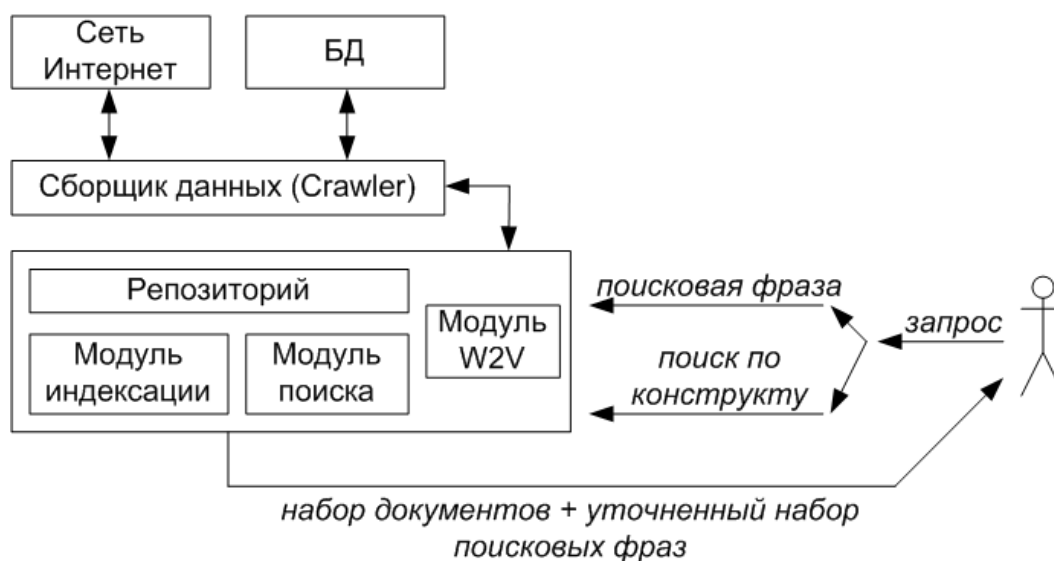


Рис. 3. Архитектура системы поиска научной информации в рамках предлагаемой методики

Предлагаемая ИПС работает в несколько этапов: предварительный анализ информации, формирующий выборку документов-кандидатов; поиск среди документов-кандидатов. Назначением модуля W2V является генерация терминов, семантически схожих с используемыми в запросе, на основе текстов документов-кандидатов. Это нужно для того, чтобы уменьшить барьер между семантическим пространством исследователя и информацией, в которой производится поиск. Кроме того, дополнительными возможностями системы является поиск по компонентам научных работ (цель, объект, предмет и прочее), таких как диссертации и авторефераты.

Алгоритм работы предлагаемой ИПС:

- *Шаг 1.* Предварительный поиск научной информации по заданному запросу, и составление предварительной выборки документов;

- *Шаг 2.* Формирование индексной информации для поиска в полученной выборке документов;

- *Шаг 3.* Обработка запроса пользователя модулем W2V для формирования набора запросов;

- *Шаг 4.* Выполнение поиска научной информации в предварительной выборке документов;

- *Шаг 5.* Выдача результатов.

Если поиск ведется в авторефератах и диссертациях, то есть в тех документах для которых на данный момент в системе имеется объект формально описывающий их структуру, то больший приоритет получают документы, имеющие совпадения с запросом в конструктах.

На текущий момент прорабатываются способы оценки точности результатов выдаваемых предлагаемой ИПС.

Список литературы / References

1. *Bornmann L., Mutz R.* Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references // Journal of the Association for Information Science and Technology, 2012. № 02.
2. *Vesset D. et al.* IDC FutureScape: Worldwide Big Data and Analytics, Predictions // [Электронный ресурс]. Режим доступа: www.cloudera.com/ (дата обращения: 20.05.2017).