

САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

На правах рукописи

Губин Максим Вадимович

Модели и методы представления текстового  
документа в системах информационного поиска

05.13.11 – Математическое и программное обеспечение  
вычислительных машин, комплексов и компьютерных сетей

АВТОРЕФЕРАТ

диссертации на соискание ученой степени  
кандидата физико-математических наук

Санкт-Петербург  
2005

Работа выполнена на кафедре информатики математико-механического факультета Санкт-Петербургского Государственного Университета.

Научный руководитель: доктор физико-математических наук,  
профессор Новиков Борис Асенович

Официальные оппоненты:

доктор физико-математических наук,  
профессор Тузов Виталий Алексеевич  
кандидат физико-математических наук,  
доцент Капустин Виктор Андреевич

Ведущая организация: Научно-исследовательский  
вычислительный центр Московского  
государственного университета им.  
М.В.Ломоносова (НИВЦ МГУ)

Защита диссертации состоится “\_\_\_” \_\_\_\_\_ 2005 года  
в \_\_\_\_\_ часов на заседании диссертационного совета Д 212.232.51  
по защите диссертаций на соискание ученой степени доктора наук  
при Санкт-Петербургском государственном университете по адресу  
198504, Санкт-Петербург, Старый Петергоф, Университетский  
пр.,28, Математико-механический факультет.

С диссертацией можно ознакомиться в Научной библиотеке  
Санкт-Петербургского государственного университета по адресу:  
199034, Санкт-Петербург, Университетская наб. 7/9.

Автореферат разослан “ ” \_\_\_\_\_ 2005 года.

Ученый секретарь  
диссертационного совета  
доктор физико-математических наук

Б. К. Мартыненко

# Общая характеристика работы

## Актуальность темы

В течение последних десятилетий наблюдается постоянно ускоряющийся рост объемов текстовой информации, хранящейся в виде электронных документов. Для эффективной работы с ними требуются современные инструменты, важную роль среди которых играют различные средства информационного поиска. Основная цель информационного поиска - помочь пользователю найти информацию, в которой он заинтересован. Система отбирает из всего имеющегося множества информации подмножество, удовлетворяющее пользователя.

Первые системы, реализующие информационный поиск, были созданы в 1950-е годы. По мере развития информационных систем размеры и количество хранящихся документов постоянно росли. Помимо самого текста, стали хранить параметры его форматирования, гипертекстовые связи между документами и другую информацию. Так, объем первых коллекций конференции TREC, посвященной вопросам информационного поиска, составлял единицы мегабайт, в то время как сейчас используются даже терабайтные коллекции [Har]. Возрастающие требования к качеству информационного поиска привели к тому, что разработчики систем стали использовать более сложные модели документа, пытаясь максимально использовать имеющиеся о нем данные [JWR00, Tak00, AvdWKvB00, Mon00, Dan92, Kar, VR79, WAW01, FC01, KZ01, SOK94, SAB93, SM97]. Текст стал рассматриваться не просто как набор терминов, а как объект с более сложной структурой, стали анализироваться и учитываться при поиске элементы форматирования текста, его структура. Развитие сети Интернет привело к появлению стандартного представления гипертекста в виде формата HTML, и сделало возможным эффективное использование при поиске анализа связей между документами [BP98, KHM03, KB03, Nav02, Hen00, Kor01, WVS+96].

Все развитие современных методов информационного поиска

можно представить как постоянное совершенствование и усложнение модели документа и методов ее использования. Актуальной темой является исследование влияния выбора модели и ее различных характеристик на качество информационного поиска.

Несмотря на высокое быстродействие современных компьютеров, невозможно обеспечить обработку запроса за приемлемое время для больших коллекций документов без использования специальных индексных структур. Актуальным вопросом является обеспечение заданных параметров эффективности, для чего постоянно предлагаются новые варианты индексных структур, алгоритмы их построения и использования [Buc85, BWZ02, KABL96, Jam00, WZA99, BP98].

Большое внимание современных исследователей привлекает проблема обеспечения компактного представления индексных структур [BWZ01a, SWYZ02, BWZ01b, Tro03, WZ99, SG]. Алгоритмы сжатия позволяют устранить избыточность и достичь заданную эффективность поиска с точки зрения требований к дисковому пространству и быстродействию.

Актуальной является задача поддержки работы с документами, текст которых может меняться в течении времени [KABL96, VV94, LWPea]. Это связано с динамическим характером многих коллекций, изменяющихся во времени, таких как Web-страницы, ленты новостей, хранилища систем документооборота.

## **Цели работы**

Исследовать модели и методы представления текстового документа в системах информационного поиска, которые учитывают взаимное положение слов. Изучить их влияние на качество информационного поиска и способы их эффективной реализации с использованием индексных структур.

## **Общая методика**

Исследования проводились в рамках классической задачи информационного поиска. Для каждой из рассматриваемых моделей выдвигалась и обосновывалась гипотеза об эффективности,

которая потом проверялась с помощью эксперимента. Для проверки использовались стандартные тестовые наборы данных РОМИП [rom].

Для анализа индексных структур и алгоритмов использовалась методика, при которой показывалась эквивалентность с заданными допущениями, структурам и алгоритмам, для которых есть общепринятые методы анализа. Результаты в последующем проверялись с помощью экспериментальных исследований.

## **Основные результаты**

В работе получены следующие основные результаты:

1. разработан вариант реализации функции взвешивания документов при информационном поиске, учитывающий взаимное положение слов, который позволил значительно увеличить качество информационного поиска;
2. предложена оригинальная реализация индексных структур, позволяющая эффективно выполнять информационный поиск с использованием описанных моделей, учитывающих взаимное положение слов;
3. разработан новый алгоритм сжатия индексной информации, позволяющий уменьшить ее объем и ускорить обработку запросов за счет уменьшения количества операций ввода-вывода;
4. предложен оригинальный алгоритм разбиения документа при индексации, который минимизирует изменение индекса при создании новой редакции документа;
5. реализован прототип системы, осуществляющий информационный поиск с использованием описываемых моделей документов;
6. проведена серия экспериментов по оценке качества поиска системы с использованием собственных методик и методики

Российского семинара по Оценке Методов Информационного Поиска (РОМИП);

7. проведена серия экспериментов по оценке эффективности сжатия индексной информации и объемам изменения индекса при создании новых редакций документов;
8. показано, что использование при информационном поиске моделей документа, учитывающих взаимное положение слов, улучшает качество поиска;
9. показано, что поиск с использованием предлагаемых моделей документов может быть эффективно реализован только с использованием специальных индексных структур.

### **Научная новизна**

Все основные научные результаты диссертации являются новыми.

### **Практическая и теоретическая ценность**

Полученные результаты могут служить теоретическим обоснованием выбора модели документа и индексной структуры для классической задачи информационного поиска. Они так же могут быть использованы как отправная точка для исследований в области других известных задач.

На практике исследованные модели и индексные структуры могут быть использованы для создания различных информационно-поисковых систем.

### **Апробация работы**

Результаты диссертации докладывались на конференциях по электронным библиотекам (RCDL'2002, Дубна, Россия, RC DL'2003, Санкт-Петербург, Россия и RC DL'2005, Пущино, Россия), Российский семинар по оценке методов информационного поиска (РОМИП'2003, Санкт-Петербург, Россия и РОМИП'2004, Пущино, Россия). Рассмотренные модели и методы представления текстового документа в системах информационного поиска были

затем использованы при разработке информационно-поисковой системы „Кодекс“.

## **Публикации**

Основные результаты диссертации изложены в шести работах [1, 2, 4, 3, 6, 5], перечисленных в конце автореферата.

## **Структура и объем диссертации**

Диссертация состоит из 5 глав со сквозной нумерацией разделов, рисунков и таблиц. Текст диссертации изложен на 95 страницах. Список литературы содержит 68 наименований.

## **Содержание работы**

В **первой главе** кратко охарактеризован общий контекст исследований. В частности, перечислены и кратко описаны основные задачи информационного поиска. Описываются подходы и методы оценки качества информационного поиска. Формулируются определение модели документа в системе информационного поиска и задачи проведенного исследования.

Во **второй главе** приводится обзор известных моделей документов в системах информационного поиска. В ней делается вывод, что первая, и самая простая, модель множества слов (bag of words) уступила свое место более сложным, но позволяющим достичь повышения качества информационного поиска.

Наиболее распространенными и известными моделями, дающими большое увеличение качества при относительно малых накладных расходах и простоте реализации являются:

1. документ, как множество весов терминов;
2. документ, как множество фрагментов;
3. документ, как узел гипертекстового графа.

В **третьей главе** рассматриваются вопросы выбора и реализации модели документа, которая учитывает взаимное

положение слов. В данной работе исследовались две модели, которые показали наиболее перспективными:

- выделение „контактных“ пар слов;
- модель с разбиением документа, при которой оценка релевантности документа производится с помощью функции, вычисляемой по „скользящему“ окну.

Каждая рассматривается как гипотеза, которая в последующем будет экспериментально проверена. В данной главе приведены обоснования выбора и описаны особенности реализации. Формулируются требования к индексной структуре, которая должна использоваться для рассматриваемых моделей. Описывается алгоритм, использующий индексную структуру, и показывается, что результат, полученный с помощью этого алгоритма, совпадает с результатом, полученным при анализе текстов документов без использования индексных структур.

В **четвертой главе** рассмотрены вопросы выбора и организации индексной структуры, позволяющей эффективно реализовывать модели, рассмотренные в предыдущей главе. Формулируются и в дальнейшем анализируются требования эффективности. Кратко описываются известные в настоящее время индексные структуры и обосновывается выбор инвертированного файла в качестве такой структуры.

Рассматриваются подходы к реализации инвертированного файла и указываются преимущества его реализации с использованием В+ дерева.

Показывается необходимость использования алгоритмов сжатия информации в инвертированном файле. Приводится описание известных алгоритмов сжатия и предлагается собственный алгоритм, обеспечивающий лучшее сжатие коротких листов вхождений для редких слов.

Формулируется проблема индексирования документов, которые изменяются во времени. Предлагается решение этой проблемы, основанное на допущении, что изменения обычно затрагивают

только отдельные участки документа. Документ при индексации разбивается на фрагменты с помощью специальных отметок, и положение слов указывается относительно этих отметок. Предлагается новый алгоритм формирования точек разбиения, основанный на хэш-функции от скользящего по тексту окна.

Показывается, что предлагаемая индексная структура имеет характеристики эффективности удовлетворяющие требованиям, сформулированным в данной главе.

В **пятой главе** представлены результаты экспериментов, выполненных с целью проверки эффективности методов и алгоритмов, описанных в предыдущих разделах этой работы. Для каждой группы экспериментов описываются коллекции документов, на которых выполнялись эксперименты, постановка эксперимента и полученные результаты.

Для оценки качества информационного поиска использовались собственные методики и методики Российского семинара по Оценке Методов Информационного Поиска (РОМИП). На основании этих методик было показано, что модель, основанная на „парах слов“ не позволяет добиться устойчивого улучшения качества поиска. Модель, основанная на использовании „скользящего окна“ показала устойчиво хорошие результаты. На рисунках 1 и 2 приведены эти результаты в виде 11-точечных графиков точность/полнота, построенных по методике ТРЕС, для коллекций РОМИП.

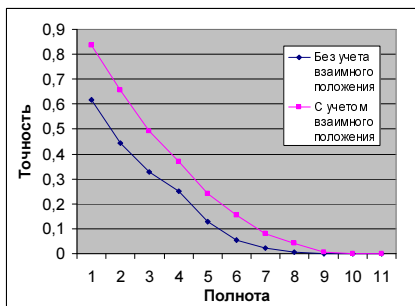


Рис. 1: Результаты web-коллекции

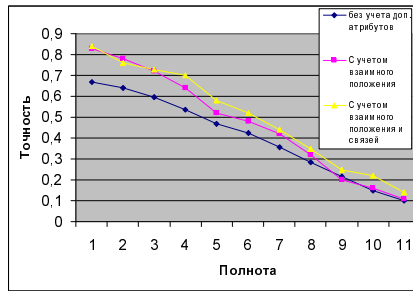


Рис. 2: Результаты legal-коллекции

Эксперименты по оценке степени сжатия индексов показали, что предлагаемый алгоритм сжатия пост-листов позволяет добиться уменьшения требуемых объемов памяти в 2 раза по сравнению с другими известными на настоящий момент методами.

Экспериментальное исследование реализации инвертированного файла, когда листы вхождения слов содержат позиции в текстах относительно специальных отметок, устойчивых к изменениям версий показало, что это позволяет значительно уменьшить объем изменяемой индексной информации при создании новой версии текста документа. При этом лучшие результаты достигнуты при формировании отметок с помощью предложенного метода с использованием хэш-функции от скользящего по тексту окна.

## Литература

- [Кор01] Некрестьянов И. Корявко А. Методы предварительной обработки данных для алгоритма клейнберга. In *Труды четвертой всероссийской конференция RCDL'2002*, volume 2, pages 215–231, 2001.
- [AvdWKvB00] A. Arampatzis, T. van der Weide, C. Koster, and P. van Bommel. Linguistically motivated information retrieval. 69, December 2000. To appear. Current-

ly available on-line from <http://www.cs.kun.nl/~avgerino/encyclopTR.ps.Z>.

- [BP98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107-117, 1998.
- [Buc85] Chris Buckley. Implementation of the smart information retrieval system. Technical report, 1985.
- [BWZ01a] D. Bahle, H. E. Williams, and J. Zobel. Compaction techniques for nextword indexes. In *Proceedings of the SPIRE Conference on String Processing and Information Retrieval*, pages 33-45, San Rafael, Chile, 2001.
- [BWZ01b] D. Bahle, H.E. Williams, and J. Zobel. Compaction techniques for nextword indexes, November 2001.
- [BWZ02] D. Bahle, H. E. Williams, and J. Zobel. Efficient phrase querying with an auxiliary index. In K. Jarvelin, M. Beaulieu, R. Baeza-Yates, and S. H. Myaeng, editors, *Proceedings of the ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 215-221, Tampere, Finland, August 2002.
- [Dan92] James A. Danowski. Wordij: A word-pair approach to information retrieval. In *TREC*, pages 131-136, 1992.
- [FC01] Massimo Melucci Franco Crivellari. Web document retrieval using passage retrieval, connectivity information, and automatic link weighting. In *The Tenth Text REtrieval Conference (TREC 2001)*, pages 624-633, 2001.
- [Har] Donna Harman. What we have learned, and not learned, from trec. In *Proc. of the BCS IRSG'2000*, pages 2-20.
- [Hav02] T. Haveliwala. Topic-sensitive pagerank. In *Proceedings of the Eleventh International World Wide Web Conference*, 2002.

- [Hen00] Monika Henzinger. Link analysis in web information retrieval. *IEEE Data Engineering. Bulletin*, 23(3):3–8, 2000.
- [Jam00] Allison Powell James. The impact of database selection on distributed searching. In *Proc. of the SIGIR'00*, 2000.
- [JWR00] K. Sparck Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments. *Inf. Process. Manage.*, 36(6):779–808, 2000.
- [KABL96] T. Koch, A. Ardö, A. Bremmer, and S. Lundberg. The building and maintenance of robot based internet search services: A review of current indexing and data collection methods. Technical report, Lund University Library, Sweden, 1996.
- [Kar] Jussi Karlgren. The basics of information retrieval: Statistics and linguistics.
- [KB03] George A. Mihaila Krishna Bharat. Hilltop: A search engine based on expert documents. <http://www.cs.toronto.edu/~georgem/hilltop/>, 2003.
- [KHMG03] Sepandar D. Kamvar, Taher H. Haveliwala, Christopher D. Manning, and Gene H. Golub. Extrapolation methods for accelerating pagerank computations. In *Proceedings of the twelfth international conference on World Wide Web*, pages 261–270. ACM Press, 2003.
- [KZ01] Marcin Kaszkiel and Justin Zobel. Effective ranking with arbitrary passages. *Journal of the American Society of Information Science*, 52(4):344–364, 2001.
- [LWPea] Lipyew Lim, Min Wang, Sriram Padmanabhan, and et al. Dynamic maintenance of web indexes using landmarks.

- [Mon00] Christof Monz. Computational semantics and information retrieval. In *Proceedings of the 2nd Workshop on Inference in Computational Semantics (ICoS-2)*, pages 1–5, 2000.
- [rom] Российский Семинар по Оценке Методов Информационного Поиска. <http://romip.narod.ru>.
- [SAB93] G. Salton, J. Allan, and C. Buckley. Approaches to Passage Retrieval in Full Text Information Systems. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 49–58, 1993.
- [SG] Satya Mahesh Rachakonda Susan Gauch, Jianying Wang. A corpus analysis approach for automatic query expansion and its extension to multiple databases.
- [SM97] M. Cutler Y. Shih and W. Meng. Using the structure of html documents to improve retrieval. In *USENIX symposium on Internet Technologies and Systems (NISTS'97)*, pages 241–251, 1997.
- [SOK94] Alan F. Smeaton, Ruairi O'Donnell, and Fergus Kellely. Indexing structures derived from syntax in TREC-3: System description. pages 100–110, 1994.
- [SWYZ02] F. Scholer, H. Williams, J. Yiannis, and J. Zobel. Compression of inverted indexes for fast query evaluation, 2002.
- [Tak00] T. Takaki. Ntt data: Overview of system approach at trec-8 ad-hoc and question answering. In *Proc. of the 8'th Text REtrieval Conference*, 2000.
- [Tro03] Andrew Trotman. Compressing inverted files. *Inf. Retr.*, 6(1):5–19, 2003.

- [VR79] C. J. Van Rijsbergen. *Information Retrieval, 2nd edition*. Butterworths, London, 1979.
- [VV94] Peter J. Varman and Rakesh M. Verma. Optimal storage and access to multiversion data. pages 1605–1620, 1994.
- [WAW01] Martin Ann Houston W. A. Woods, Stephen Green Paul. Aggressive morphology and lexical relations for query expansion. Technical report, 2001.
- [WVS<sup>+</sup>96] Ron Weiss, Bienvenido Velez, Mark Sheldon, Chanathip. Nemprempre, Peter Szilagyi, Andrzej Duda, and David Gifford. HyPursuit: A hierarchical network search engine that exploits content-link hypertext clustering. In *Proc. of Seventh ACM Conference on Hypertext*, March 1996.
- [WZ99] Hugh E. Williams and Justin Zobel. Compressing integers for fast file access. *The Computer Journal*, 42(3):193–201, 1999.
- [WZA99] H. E. Williams, J. Zobel, and P. Anderson. What’s next? Index structures for efficient phrase querying. In J. Roddick, editor, *Proceedings of the Australasian Database Conference*, pages 141–152, Auckland, New Zealand, 1999.

## Работы автора по теме диссертации

- [1] Губин М.В. Изучение статистики встречаемости терминов и пар терминов в текстах для выбора методов сжатия инвертированного файла. In *Труды RCDL-2002*, volume 2, pages 26–38, 2002.
- [2] Губин М.В. Исследование качества информационного поиска с использованием пар слов. In *Труды RCDL-2003*, pages 186–191, 2003.

- [3] Губин М.В. Опыт участия ИС „Кодекс“ в РОМИП 2003. In *Труды РОМИП'2003*, pages 31–42, 2003.
- [4] Губин М.В. Электронная библиотека многоверсионных текстовых документов. In *Труды RCDL-2004*, pages 169–174, 2004.
- [5] Губин М.В. Модели и методы представления текстовых документов в системах информационного поиска. *Научно-Техническая Информация*, 1(12):12–23, 2004.
- [6] Губин М.В. Участие ИПС „Кодекс“ в семинаре РОМИП 2004. In *Труды РОМИП'2004*, pages 28–39, 2004.

Подписано к печати 18.03.2005.

Бумага офсетная. Печать ризографическая.

Объем 1 усл. п.л. Тираж 100 экз. Заказ 3401.

Отпечатано в ООО "Акрон" с оригинал-макета заказчика.

191186, Санкт-Петербург, ул. Малая Морская, д. 26, тел./факс: 312-6500

Бесплатно