УДК 004.912

РАСПРЕДЕЛЕННАЯ ИНСТРУМЕНТАЛЬНАЯ СРЕДА СЛОВАРНОГО МОРФОЛОГИЧЕСКОГО АНАЛИЗА ДЛЯ ОБРАБОТКИ РУССКОГО ЯЗЫКА

Д.А. Усталов, М.Л. Гольдштейн

В статье рассмотрен подход к масштабированию сервиса морфологического разбора слов естественного языка при обработке различных коллекций документов на русском языке. Выполнен обзор и критический анализ существующих решений. Сформированы требования к инструментальной среде словарного морфологического анализатора. Распределенная архитектура Web–сервиса морфологического анализа, предназначенного для обработки крупных коллекций документов на русском языке, представлена в виде структурной модели. Данная архитектура реализована в виде прототипа системы на языке программирования Ruby. Приведена структура используемого морфологического словаря в виде реляционной схемы. Испытания данного метода в распределенной вычислительной среде показали линейную масштабируемость предлагаемого решения. Конфигурация эксперимента включает систему генерации нагрузки в виде HTTP–запросов, систему балансировки нагрузки на рабочие узлы распределенной системы, серверы приложений с функционирующим анализатором и базу данных морфологического словаря, а также кэширующий узел для снижения издержек при выполнении запросов к словарю. Применение данного подхода позволяет получить линейный рост производительности в распределенных системах автоматической обработки больших объемов текста.

Ключевые слова: распределенные вычисления, обработка естественного языка, корпусная лингвистика, обработка больших объемов данных, морфологический анализ.

Введение

Задачи, связанные с обработкой естественного языка актуальны как для научной, так и для практической деятельности человека. На сегодняшний день существует большое разнообразие инструментария для подобных целей, но это, в основном, касается зарубежных языков. Русскоязычный же инструментарий достаточно скуп. Работы по развитию такого инструментария сегодня ведутся в рамках ряда программ, в т. ч. и программы фундаментальных исследований Президиума РАН «Корпусная лингвистика» [1]. Одно из актуальнейших научных направлений этой программы — обработка крупных коллекций документов — корпусов текстов, с помощью современного инструментария. Важным разделом этого направления является морфологическая разметка корпусов текстов. При этом объем данных для такой работы может измеряться гигабайтами, и разметка таких коллекций может быть выполнена только в рамках технологий облачных вычислений и систем автоматической обработки текста, например, *GATE Cloud* [2].

Именно распределенные масштабируемые инструментальные среды обработки текста, в т. ч. распределенный морфологический анализатор, могут быть востребованы в различных отраслях. Например, в рамках концепции «образование через всю жизнь» аудит знаний нашел широкое применение в современных системах менеджмента качества образования [3]. Наличие в таких системах подсистем морфологического, синтаксического и семантического анализов, может способствовать повышению культуры речи и письма выпускника вуза путем сканирования и анализа различных письменных работ студента на протяжении всего периода обучения.

1. Аналогичные работы

На сегодняшний день существует ряд популярных решений для морфологического анализа русскоязычных текстов: *mystem* [4], *Snowball* [5] и *stemka* [6].

Вышеобозначенные решения предназначены для работы в локальных окружениях, и не имеют встроенных средств для построения распределенных систем морфологического анализа, которые необходимы для эффективной обработки больших объемов данных. Данный недостаток может быть преодолен путем использования таких инструментов распределения задач, как *Gearman* [7].

Другим ограничением данных решений является отсутствие унифицированного формата представления грамматических характеристик, что несколько затрудняет как применение этих анализаторов в сторонних системах, так и обмен результатами анализа между приложениями.

В рамках проекта *MULTEXT–East* [8] составляется консолидированный набор морфосинтаксических дескрипторов для восточноевропейских и ряда западноевропейских языков.

Использование достижений проекта *MULTEXT–East* позволяет унифицировать формат представления грамматических характеристик и сформировать единый программный интерфейс для удобной работы с результатами морфологического разбора, устранив данное ограничение.

2. Морфологический анализатор Myaso

Морфологический анализатор общего назначения *Myaso* (название образовано как акроним от фразы «*my analysis system is open*») разрабатывается с 2010 года в целях апробации новаторских методов морфологической обработки текста на русском и английском языках.

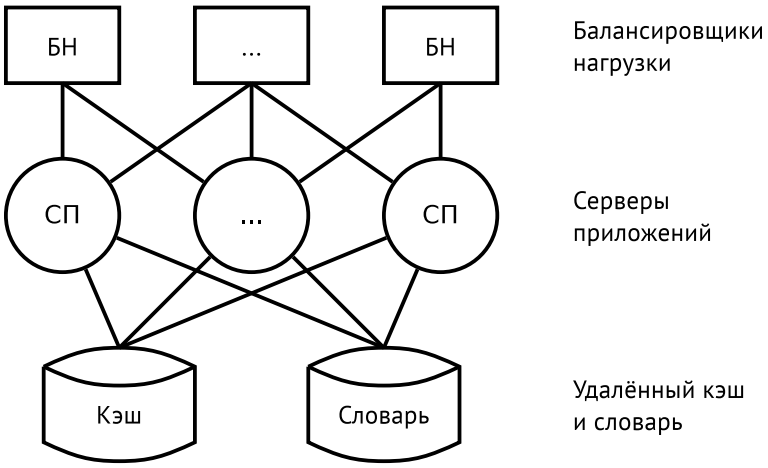

Рис. 1. Структура Web–сервиса морфологического разбора текста

Анализатор *Myaso* доступен как в виде встраиваемого решения[1], так и в виде Web–сервиса [9], спроектированного по архитектуре *REST*.

Структурная схема Web–сервиса представлена на рис. 1:

1. HTTP-запрос с данными для обработки приходит на фронтэнд — один из балансировщиков нагрузки (БН);

[1] *Ruby Gem* — подключаемая библиотека для среды выполнения *Ruby*.

2. каждый балансировщик нагрузки хранит в себе сведения о нагруженности бэкэндов — серверов приложений (СП) и перенаправляет запрос к наиболее доступному в данный момент серверу приложений, способному выполнить обработку исходного запроса;

3. сервер приложений проверяет наличие результатов анализа запрошенного слова в кэше, и при его наличии, возвращает ответ на запрос клиента, после чего закрывает соединение;

4. при отсутствии результатов разбора слова в кэше, сервер приложений выполняет анализ запрошенного слова, используя данные из морфологического словаря, сохраняет ответ в кэше, возвращает полученное значение клиенту и закрывает соединение.

Структура морфологического словаря (лексикона) анализатора *Myaso* основана на морфологическом словаре проекта ДИАЛИНГ [10]. В данном случае, словарь представлен в виде реляционной схемы на рис. 2 и ориентирован на хранение грамматической информации в формате морфосинтаксических описателей *MULTEXT–East* [8].

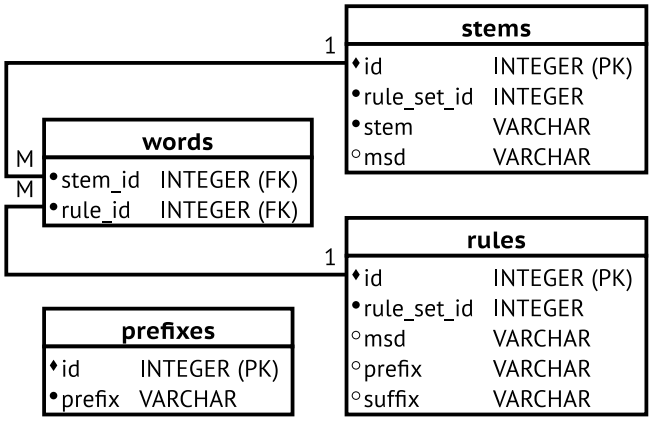

Рис. 2. Структура морфологического словаря анализатора *Myaso*

Принцип работы анализатора основан на работе [11], однако в рассматриваемом прототипе анализатора *Myaso* не реализован механизм синтеза морфологических гипотез для слов, неизвестных словарю.

3. Экспериментальная проверка

Для оценки эффективности предложенного решения выполнено несколько запусков анализатора *Myaso* в различных режимах и конфигурациях.

Анализатор *Myaso* был запущен на узлах `one1`, `one2` и `one3` в режиме Web–сервиса. Балансировка нагрузки между узлами осуществлялась реверсным прокси-сервером *haproxy* [12], установленном на узле `cf`. Поведение *Myaso* исследовалось в двух режимах: как с подключением кэш–сервера `one4`, так и без него.

Эксперимент выполнялся в вычислительном центре ИММ УрО РАН при следующих настройках (рис. 3):

1. узел `cf` с балансировщиком нагрузки, один рабочий узел `one1`;

2. узел `cf` с балансировщиком нагрузки, два рабочих узла: `one1`, `one2`;

3. узел `cf` с балансировщиком нагрузки, три рабочих узла: `one1`, `one2`, `one3`.

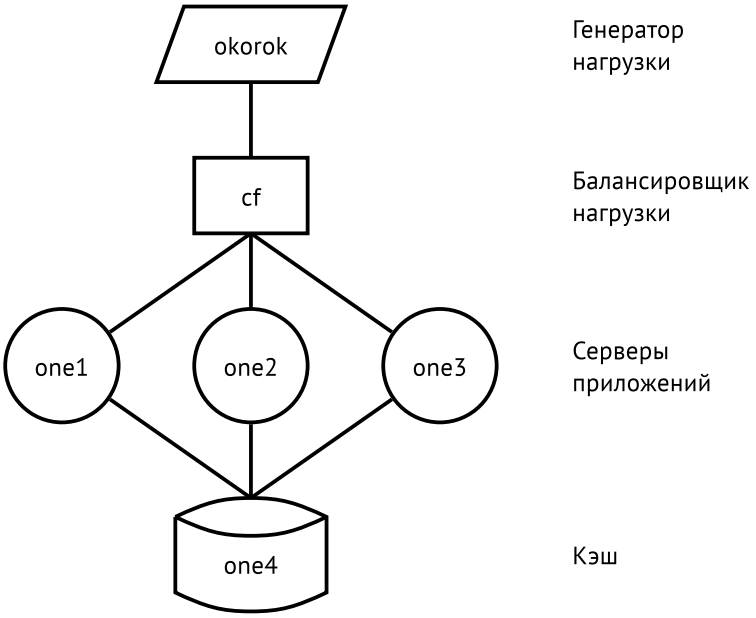

Рис. 3. Конфигурация эксперимента

Сравнение проводилось путем обработки коллекции документов [13], состоящей из 202 документов, содержащих 28464 русскоязычных лексем. Средний объем документа в данной коллекции составляет 140, 91 лексем.

Перед каждым запуском тестов анализатора *Myaso* с подключенным кэшем, выполнялось обнуление кэша. Словарь анализатора *Myaso* хранится в файловой нереляционной базе данных *Tokyo Cabinet* [14].

3.1. Конфигурация узлов

Балансировщик нагрузки расположен на узле `cf`, а генератор нагрузки – на узле `okorok`. Оба узла представляют из себя виртуальные машины эквивалентной конфигурации с операционной системой *CentOS* 6.2 (x86_64), установленные в гипервизоре *VMware ESXi*. Сервер *HP ProLiant DL165 G7 6172*, на котором функционирует гипервизор, имеет следующую конфигурацию:

- два 12-ядерных процессора *AMD Opteron^{TM} 6172* @ 2,1 ГГц (виртуальным машинам доступно по 3 процессорных ядра);
- 16 ГБ оперативной памяти (виртуальным машинам доступно по 4 ГБ);
- каждой виртуальной машине предоставлено 10 ГБ дискового пространства;
- сетевая карта *HP NC362i Integrated Dual Port Gigabit Server Adapter*.

Рабочие узлы `one1`, `one2`, `one3` и узел кэша `one4` работают под операционной системой *Scientific Linux 6.1 (x86_64)*, установленной на четырех полноценных физических серверах *Fujitsu-Siemens Computers PRIMERGY RX330 S1* следующей конфигурации:

- два 2-ядерных процессора *AMD Opteron^{TM} 2218* @ 2,6 ГГц;
- 8 ГБ оперативной памяти;

- жесткий диск *Seagate Barracuda ES 250GB Serial ATA II 7200RPM 16MB*;
- сетевая карта *Broadcom BCM5715 Gigabit Ethernet*.

Сервису *memcached* [15] на узле `one4` предоставлено 6 ГБ оперативной памяти для работы.

Все перечисленные узлы соединены по технологии *Gigabit Ethernet*.

3.2. Результаты

Результаты эксперимента приведены в таблице и изображены на рис. 4.

Значения P_1, P_2 и P_3, соотвествующие результатам эксперимента, вычислены по формуле 1:

$$P_n = \frac{wc}{t}, \tag{1}$$

где n — количество рабочих узлов;
wc — общее количество слов в корпусе текстов (использованная коллекция документов содержит 28464 русскоязычных лексемы);
t — суммарное время обработки всех слов корпуса текстов данным анализатором на n рабочих узлах.

Таблица

Результаты сравнения режимов работы анализатора *Myaso*

Анализатор	$\mathbf{t_1}$	$\mathbf{P_1}$	$\mathbf{t_2}$	$\mathbf{P_2}$	$\mathbf{t_3}$	$\mathbf{P_3}$
Myaso (с кэшем)	1905.40	14.94	963.34	29.55	657.92	43.26
Myaso (без кэша)	2076.13	13.71	1047.39	27.18	716.01	39.75

3.3. Анализ

Исходя из результатов, приведенных в табице, можно отметить, что анализатор *Myaso* имеет неэффективную реализацию подсистемы кэширования, из-за чего практически не наблюдается прироста производительности при использовании кэша.

Анализатор *Myaso* предназначен для работы с применением словарей и является морфологическим анализатором общего назначения, т. е. помимо операции стемминга способен решать задачи, свойственные другим анализаторам на основе словаря: производить склонение слова в рамках парадигмы, определять грамматические характеристики слова, и т. д.

4. Заключение

В работе представлен прототип распределенной системы морфологического разбора слов русского языка и проведена его сравнительная оценка с аналогичными решениями.

Результаты проведенного эксперимента диктуют предпосылки к дальнейшей работе:

1. язык программирования *Ruby* значительно жертвует производительностью приложения в обмен на скорость разработки, поэтому актуален вопрос частичного или полного переписывания анализатора на языке, способном компилироваться в «родной» код операционной системы;

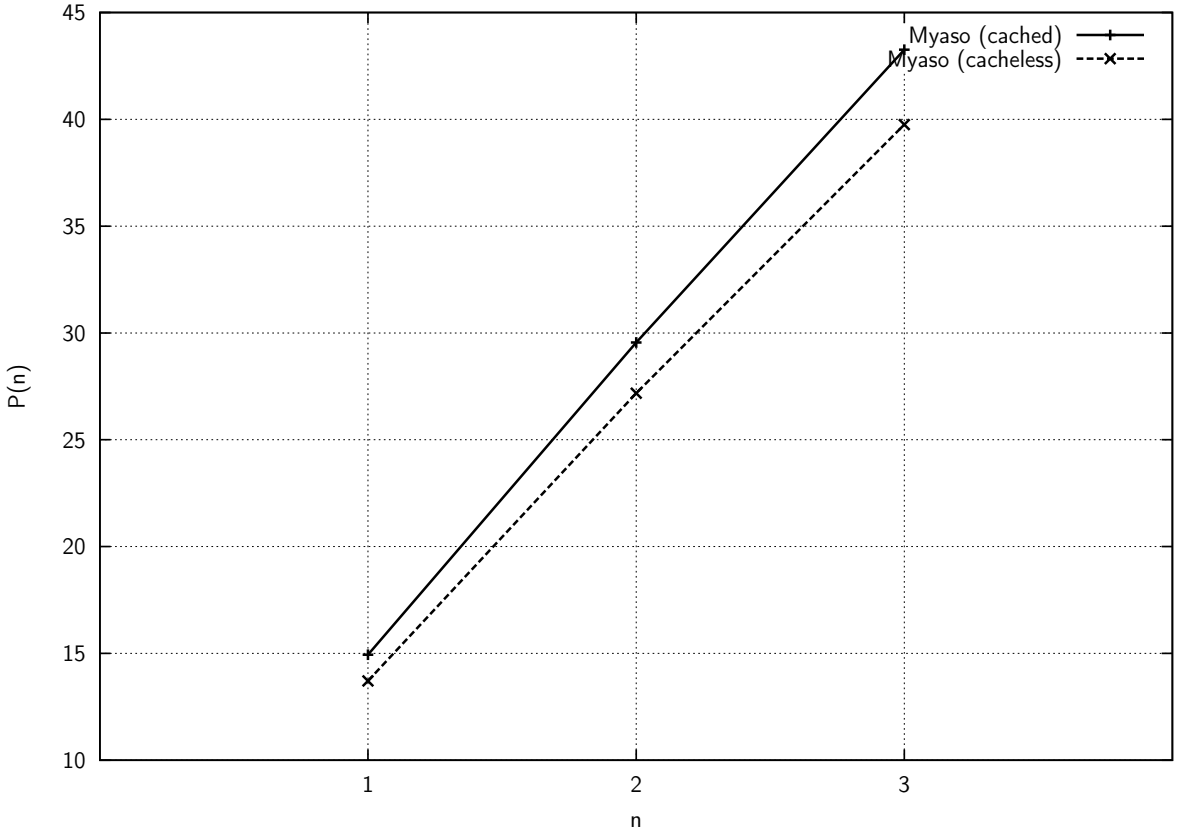

Рис. 4. Результаты сравнения режимов работы анализатора *Myaso*

2. применение реляционной СУБД (например, *PostgreSQL*) для хранения морфологического словаря позволит повысить производительность работы с лексиконом, благодаря его реляционной структуре;

3. рассмотреть возможность модификикации подсистемы кэширования данных с целью повышения производительности;

4. необходимо реализовать систему формирования гипотез для слов, неизвестных морфологическому словарю (аналогичную описанной в работе [11]).

Среди преимуществ представленного анализатора *Myaso* можно отметить:

1. в отличие от ближайшего аналога [4], анализатор *Myaso* является свободным программным обеспечением и доступен любому желающему на условиях лицензии *MIT*;

2. анализатор может быть использован в качестве библиотеки языка *Ruby*, но при этом имеются встроенные механизмы построения линейно–масштабируемых Web–сервисов морфологической обработки текста;

3. анализатор *Myaso* построен на основе алгоритмов [11], доказавших свою высокую эффективность в области информационного поиска [4], корпусной лингвистики [16], и др.

Репозиторий с исходным кодом морфологического анализатора *Myaso* доступен по адресу https://github.com/eveel/myaso.

Работа поддержана грантом УрО РАН №РПЦ-12-П10 и грантом конкурса фундаментальных исследований, выполняемых совместно с организациями СО и ДВО РАН, государственных академий наук России, национальных академий наук стран СНГ и отраслевых академий и финансируемых из средств Уральского отделения РАН №12-С-1-1012.

Литература

1. Корпусная лингвистика. [Электронный ресурс] // [сайт].
   URL: http://www.corpling-ran.ru/index.html (дата обращения 20.05.2012).
2. GATE Cloud — a New Way to Mine the Web. [Электронный ресурс] // [сайт]. URL: http://gatecloud.net (дата обращения 20.05.2012).
3. Система менеджмента качества, оперативный контроль и анализ образовательного процесса / А.Л. Шестаков, А.И. Сидоров, Л.А. Шефер, Е.В. Гичкина // Вестн. Ленинград. гос. ун-та имени А.С. Пушкина. – 2009. – № 1. – С. 177–194.
4. mystem [Электронный ресурс] // [сайт].
   URL: http://company.yandex.ru/technologies/mystem (дата обращения 20.05.2012).
5. Snowball [Электронный ресурс] // [сайт]. URL: http://snowball.tartarus.org
   (дата обращения 20.05.2012).
6. Stemka [Электронный ресурс] // [сайт]. URL: http://www.keva.ru/stemka/stemka.html (дата обращения 20.05.2012).
7. Gearman [Электронный ресурс] // [сайт]. URL: http://gearman.org
   (дата обращения 20.05.2012).
8. Erjavec, T. MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora / T. Erjavec // Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC'10. – Malta.– 2010. – С. 2544–2547.
9. Myaso [Электронный ресурс] // [сайт]. URL: http://myaso.eveel.ru
   (дата обращения 20.05.2012).
10. AOT :: Технологии [Электронный ресурс] // [сайт].
    URL: http://aot.ru/technology.html (дата обращения 20.05.2012).
11. Segalovich, I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine / I. Segalovich // Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications. MLMTA'03. – Las Vegas.– 2003. – С. 273–280.
12. HAProxy - The Reliable, High Performance TCP/HTTP Load Balancer [Электронный ресурс] // [сайт]. URL: http://haproxy.1wt.eu (дата обращения 20.05.2012).
13. Коллекция блог–записей [Электронный ресурс] // [сайт]. URL: http://plove.eveel.ru (дата обращения 20.05.2012).
14. Tokyo Cabinet: a modern implementation of DBM [Электронный ресурс] // [сайт]. URL: http://fallabs.com/tokyocabinet (дата обращения 20.05.2012).
15. Memcached – a distributed memory object caching system [Электронный ресурс] // [сайт]. URL: http://memcached.org (дата обращения 20.05.2012).
16. Программные средства. Национальный корпус русского языка [Электронный ресурс] // [сайт]. URL: http://www.ruscorpora.ru/corpora-progr.html
    (дата обращения 20.05.2012).

Дмитрий Алексеевич Усталов, бакалавр информационных систем, старший программист отдела вычислительной техники, Институт математики и механики УрО РАН (г. Екатеринбург, Российская Федерация), dau@imm.uran.ru.

Михаил Людвигович Гольдштейн, кандидат технических наук, заведующий отделом вычислительной техники, Институт математики и механики УрО РАН (г. Екатеринбург, Российская Федерация), mlg@imm.uran.ru.

MSC 68T50

A Distributed Dictionary – Based Morphological Analysis Framework for Russian Language Processing

D.A. Ustalov, Institute of Mathematics and Mechanics,
Ural Branch of the Russian Academy of Sciences (Yekaterinburg, Russian Federation),
M.L. Goldstein, Institute of Mathematics and Mechanics,
Ural Branch of the Russian Academy of Sciences (Yekaterinburg, Russian Federation)

This article describes an approach to scaling service morphological parsing of words of natural language processing of various collections of documents in Russian. An overview and critical analysis of existing solutions. The requirements workbench vocabulary morphological analyzer were established. The distributed architecture of the web service morphological analysis, designed to a handle large collections of documents in Russian, presented the form of a structural model. This architecture is implemented as a prototype system in the programming language Ruby. The structure used in the morphological dictionary of a relational schema. Tests of this method in a distributed computing environment showed linear scalability of the proposed solutions. The configuration of the experiment involves the generation of the system load as a HTTP requests, system load balancing working nodes of a distributed system, application servers with a functioning database analyzer and morphological dictionary, as well as a caching node to reduce costs when you run queries to the dictionary. Applying this approach provides a linear increase in performance in distributed systems, automated processing of large volumes of text.

Keywords: distributed computing, natural language processing, corpus linguistics, data–intensive computing, morphological analysis.

References

1. *Corpus Linguistics.* Available at: http://www.corpling-ran.ru/index.html (accessed 20 May 2012).
2. *GATE Cloud — a New Way to Mine the Web.* Available at: http://gatecloud.net (accessed 20 May 2012).
3. Shestakov A.L., Sidorov A.I., Shaefer L.A., Gichkina E.V. The System of the Management Quality, Operational Control and Analysis of the Educational Process. *Vestnik Leningradskogo gosudarstvennogo universiteta imeni A.S. Pushkina* [Herald of the Leningrad State University], 2009, no. 1, pp. 177–194. (in Russian)
4. *Mystem.* Available at: http://company.yandex.ru/technologies/mystem (accessed 20 May 2012).
5. *Snowball.* Available at: http://snowball.tartarus.org (accessed 20 May 2012).
6. *Stemka.* Available at: http://www.keva.ru/stemka/stemka.html (accessed 20 May 2012).
7. *Gearman.* Available at: http://gearman.org (accessed 20 May 2012).

8. Erjavec T. MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC'10.* Malta, 2010, pp. 2544–2547.
9. *Myaso.* Available at: http://myaso.eveel.ru (accessed 20 May 2012).
10. *AOT :: Technologies.* Available at: http://aot.ru/technology.html (accessed 20 May 2012).
11. Segalovich I. A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine. *Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications. MLMTA'03.* Las Vegas, 2003, pp. 273–280.
12. *HAProxy – The Reliable, High Performance TCP/HTTP Load Balancer.* Available at: http://haproxy.1wt.eu (accessed 20 May 2012).
13. *Blog Posts Collection.* Available at: http://plove.eveel.ru (accessed 20 May 2012).
14. *Tokyo Cabinet: a Modern Implementation of DBM.* Available at: http://fallabs.com/tokyocabinet (accessed 20 May 2012).
15. *Memcached – a Distributed Memory Object Caching System.* Available at: http://memcached.org (accessed 20 May 2012).
16. *Software. Russian National Corpus Language.* Available at: http://www.ruscorpora.ru/corpora-progr.html (accessed 20 May 2012).

Поступила в редакцию 8 июня 2012 г.