https://doi.org/10.5281/zenodo.8207369
УДК 004.738.5:004.6

✉ *С. Ф. Липницкий, Л. В. Степура*

Компьютерная система интернет-мониторинга научно-технической информации

Липницкий Станислав Феликсович,
доктор технических наук, Объединенный институт проблем информатики Национальной академии наук Беларуси, одел совместных программ космических и информационных технологий, главный научный сотрудник (Минск, Беларусь)
Email: *lipn@newman.bas-net.by*

Степура Людмила Васильевна,
Объединенный институт проблем информатики Национальной академии наук Беларуси, одел совместных программ космических и информационных технологий, научный сотрудник (Минск, Беларусь)
Email: *stepura@newman.bas-net.by*

Аннотация. Представлена концепция автоматизированной системы интернет-мониторинга научно-технической информации при решении задач поддержки процессов принятия решений в различных предметных областях. Функциональными компонентами системы являются три основные подсистемы: подсистема лингвостатистического анализа текстовых документов, подсистема их лексико-семантической обработки и подсистема информационного поиска. Сформулированы задачи, решаемые данными подсистемами. Первая подсистема обеспечивает создание лингвистических словарей системы интернет-мониторинга, а также вычисление информативности слов, предложений и текстов. Второй подсистемой решаются задачи реферирования, рубрицирования и оценки тональности текстовых документов и сообщений. Третья подсистема осуществляет индексирование веб-страниц, их документальный поиск и фактографический поиск сообщений на найденных веб-страницах.

Ключевые слова: *алгоритм, индексирование, интернет-мониторинг, информативность, информационный поиск, критерий выдачи, реферирование, рубрицирование.*
Для цитирования: *Липницкий, С.Ф. Компьютерная система интернет-мониторинга научно-технической информации / С. Ф. Липницкий, Л. В. Степура // Библ.-информ. дискурс. – 2023. – Т. 3, No 1. – С. 5–10. https://doi.org/10.5281/zenodo.8207369*
Статья поступила: 16.06.2023
Статья принята в печать: 28.06.2023
Статья опубликована: 30.06.2023

✉ *Stanislav F. Lipnitsky, Ludmila V. Stepura*

Computer system for Internet monitoring of scientific and technical information

Stanislav F. Lipnitsky
Doctor of Sciences in Engineering, The United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Chief Researcher (Minsk, Belarus)
Email: *lipn@newman.bas-net.by*

Ludmila V. Stepura
The United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Researcher (Minsk, Belarus)
Email: *lipn@newman.bas-net.by*

Abstract. The concept of an automated system for Internet monitoring of scientific and technical information in solving problems of supporting decision-making processes in various subject areas is presented. The functional components of the system are three main subsystems: a subsystem for linguistic and statistical analysis of text documents, a subsystem for their lexico-semantic processing, and an information retrieval subsystem. The tasks solved by these subsystems are formulated. The first subsystem ensures the creation of linguistic dictionaries of the Internet monitoring system, as well as the calculation of the information content of words, sentences and texts. The second subsystem solves the problems of summarizing, categorizing and evaluating the tone of text documents and messages. The third subsystem performs indexing of web pages, their documentary search and factual search for messages on the found web pages.

Keywords: *algorithm, indexing, Internet monitoring, information content, information retrieval, issuance criterion, summarizing, categorization.*

For citation: *Lipnitsky S. F., Stepura L. V. Computer system for Internet monitoring of scientific and technical information. Bibliotechno-informatsionnyi diskurs = Library & Information Discourse, 2023, vol. 3, no. 1, pp. 5–10 (in Russian). https://doi.org/10.5281/zenodo.8207369*

The article was received: 16.06.2023
The article was accepted for publication: 28.06.2023
Article published: 30.06.2023

Введение

Термин «информационный мониторинг» появился в начале 1990-х годов. Под мониторингом понималась технология систематического сбора и обработки информации с целью использования ее при принятии решений в различных предметных областях [1]. Необходимость автоматизации процессов мониторинга связана с большими интеллектуальными и финансовыми затратами предприятий и организаций при «ручном» выполнении этих работ. По данным из статьи [1], поиск в Интернете необходимых сведений и их систематизация занимает в среднем 35% рабочего времени сотрудников, а 14% фонда заработной платы – это нерационально использованный бюджет. С другой стороны, как утверждается в работе [2], 80–90% необходимой для принятия решений информации может быть получено из открытых источников.

Предложенный в данной статье подход к интернет-мониторингу, в отличие от существующих, основан на использовании тематических корпусов текстов (совокупностей текстов по конкретной тематике), что обеспечивает адаптацию системы мониторинга к решаемой задаче и информационным потребностям пользователей, а также независимость программного комплекса от входных языков [3]. Такие корпусы текстов могут создаваться предварительно под прогнозируемые задачи, а также формироваться оперативно после поступления запроса (динамические корпусы текстов).

Функциональными компонентами системы информационного мониторинга Интернета являются три основные подсистемы:

– подсистема лингвостатистического анализа текстовых документов;

– подсистема лексико-семантической обработки текстовых документов;

– подсистема информационного поиска.

Лингвостатистический анализ текстовых документов

Подсистема лингвостатистического анализа текстовых документов предназначена для создания лингвистических словарей системы интернет-мониторинга, а также для вычисления информативности слов, предложений и текстов.

Корпусы текстов и лингвистические словари

Задачи создания и использования корпусов текстов решаются в рамках специального раздела языкознания – корпусной лингвистики. Под корпусом текстов понимают совокупность текстов, накопленных и размеченных по определенным принципам в зависимости от назначения. В случае отсутствия разметки эти совокупности называют корпусами текстов первого порядка. В базе данных системы интернет-мониторинга имеются тематические и полный корпусы текстов. Тематический корпус текстов – это набор неструктурированных текстовых документов по конкретной тематике. Полный корпус текстов объединяет все тематические корпусы.

В системе интернет-мониторинга созданы следующие лингвистические словари:

– частотный словарь словоформ;

– словарь словоизменительных парадигм;

– словарь синонимичных словоформ.

В частотном словаре словоформ каждой словоформе поставлены в соответствие частота в полном корпусе текстов, частоты в тематических корпусах и номер (код) парадигмы.

В словаре словоизменительных парадигм представлены словоизменительные парадигмы, используемые при вычислении информативности слов.

Словарь синонимичных словоформ состоит из групп синонимов, которые используются при определении их информативности (две синонимичные словоформы считаются двумя вхождениями лексемы в текст документа).

Словарь словоформ является единым для всех тематических разделов предметной области, т. е. в него попадают словоформы из всех тематических корпусов текстов. Формируется этот словарь программно. Имеется возможность периодического обновления словаря словоформ после добавления статей в тематический корпус текстов.

Словарь парадигм формируется и актуализируется в человеко-машинном режиме с использованием соответствующего программного инструментария эксперта-лингвиста.

Вычисление информативности слов

При вычислении информативности слов используются их абсолютные частоты в тексте (если его объем достаточно большой) или в релевантном тексту тематическом корпусе текстов (если это краткое сообщение, т. е. объем текста небольшой), а также абсолютные частоты слов в полном корпусе текстов. Информативность слова вычисляется как отношение этих частот.

При этом:

– частота слова в документе – это сумма частот всех словоформ, встречающихся в документе и являющихся словоизменениями исходной словоформы или ее синонимами, зафиксированными в словаре словоизменительных парадигм и в словаре синонимичных словоформ;

– частота слова в полном корпусе текстов – это сумма частот всех словоформ в полном корпусе текстов, являющихся также словоизменениями исходной словоформы или ее синонимами.

Вычисление информативности предложений

Рассмотрим произвольное предложение некоторого текста. Представим его вектором в евклидовом пространстве. Размерность этого пространства равна числу слов в полном корпусе текстов. При этом компоненты вектора суть значения информативности соответствующих слов предложения. Если же некоторых слов нет в предложении, то соответствующие координаты вектора равны нулю.

С учетом принятых соглашений естественно предположить, что информативность предложения – это длина его вектора. Данное допущение позволяет сравнивать информативность предложений в рамках одного и того же текста. Для сравнения же значений информативности предложений из разных текстов необходима нормализация понятия информативности. С целью нормализации рассмотрим вектор, все компоненты которого равны единице. Тогда нормализованную информативность рассматриваемого предложения можно интерпретировать как проекцию вектора данного предложения на указанный вектор с единичными координатами, т. е. скалярное произведение этих векторов, деленное на длину «единичного» вектора, которая равняется корню квадратному из размерности «единичного» вектора.

Вычисление информативности текстов

Пусть имеется некоторый текст. По аналогии с предложением представим данный текст вектором, координатами которого являются значения информативности всех его предложений. Тогда нормализованная информативность текста равна скалярному произведению вектора с единичными компонентами и вектора текста, деленному на длину

«единичного» вектора.

Пример реализации

На рисунке 1 представлен пример реализации подсистемы лингвостатистического анализа текстовых документов.

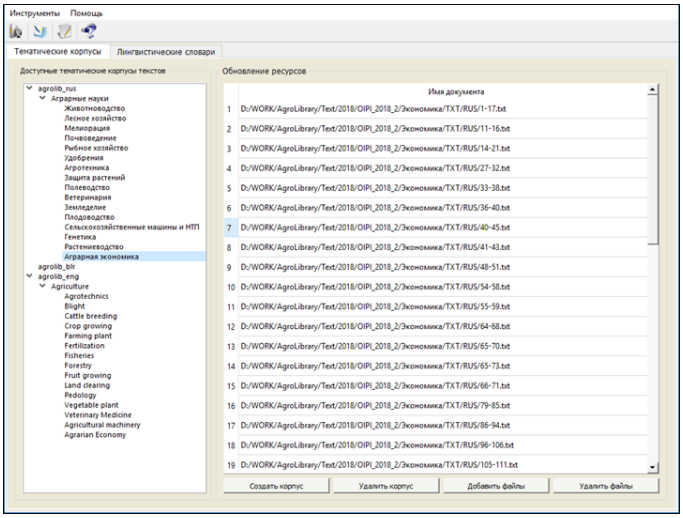


Рисунок 1. – Вкладка «Тематические корпусы» подсистемы лингвостатистического анализа текстовых документов
Figure 1. – Tab "Thematic corpora" of the subsystem of linguistic and statistical analysis of text documents

Лингвостатистический анализ текстовых документов

Посредством подсистемы лексико-семантической обработки текстовых документов решаются задачи их реферирования и рубрицирования, а также оценивается тональность текстовых сообщений.

Реферирование текстовых документов

Реферирование текстовых документов осуществляется в два этапа. На первом этапе вычисляется информативность всех предложений реферируемого текста, и предложения сортируются по убыванию значений их информативности. На втором этапе исключается «нижняя» часть полученного списка в соответствии с заданным пользователем объемом реферата. В оставшейся «верхней» части восстанавливается порядок предложений согласно их порядку в реферируемом тексте. Реферат сформирован.

Рубрицирование текстовых документов

При рубрицировании текстовых документов каждой рубрике сопоставляется релевантный тематический корпус текстов. Поиск адекватной документу рубрики реализуется путем смыслоотождествления поисковых образов рубрицируемого текста и тематических корпусов текстов. Выбирается рубрика по максимуму значения используемого критерия выдачи.

Алгоритм рубрицирования текстов работает в три шага. На первом шаге индексируется подлежащий рубрикации текстовый документ как запрос пользователя на поиск рубрики. На втором шаге корректируется поисковое предписание, полученное в результате индексирования рубрицируемого текста. При коррекции по первоначальному поисковому предписанию проводится поиск релевантных документов в полном корпусе текстов. Найденное множество документов индексируется и строится его поисковый образ, т. е. откорректированное поисковое предписание. На третьем шаге осуществляется поиск релевантных рубрик по откорректированному поисковому предписанию. Из найденных рубрик выбирается та, которой соответствует наибольшее значение критерия выдачи. К найденной рубрике относится исходный текст.

Оценка тональности текстовых сообщений

Для оценки тональности текстовых сообщений используется совокупность тонально окрашенных тематических корпусов текстов. Каждому корпусу соответствует некоторая оценка тональности. При n-балльной шкале оценок количество таких корпусов должно быть равно n. Всякий корпус включает текстовые документы одинаковой тональности. В простейшем случае формируются два корпуса текстов с тонально окрашенной лексикой. Первый корпус создается для анализа положительной тональности, а второй – для анализа отрицательной тональности. Оценка тональности текстовых сообщений реализуется путем поиска наиболее релевантных им тонально окрашенных корпусов текстов.

Пример реализации

На рисунках 2 и 3 представлен пример реализации подсистемы лексико-семантической обработки текстовых документов.

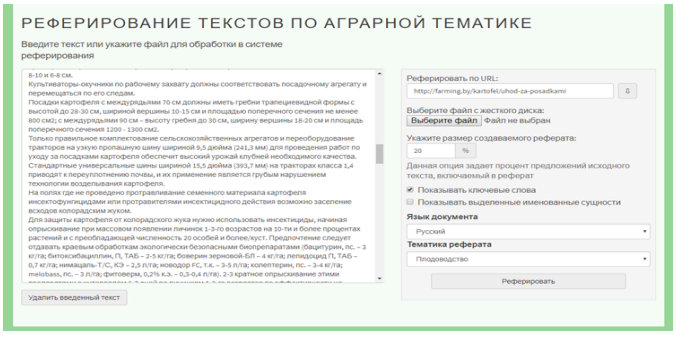
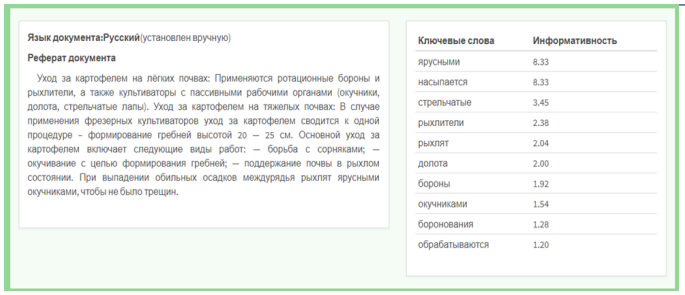


Рисунок 2. – Вкладка «Реферирование текстов» в подсистеме лексико-семантической обработки текстовых документов
Figure 2. – Tab "Text summarization" in the subsystem of lexico-semantic processing of text documents

Информационный поиск

В состав подсистемы информационного поиска входят информационно-программные средства: индексирования веб-страниц; документального поиска веб-страниц; фактографического поиска сообщений на найденных веб-страницах [4].



Язык документа:Русский(установлен вручную)

Реферат документа

Уход за картофелем на лёгких почвах: Применяются ротационные бороны и рыхлители, а также культиваторы с пассивными рабочими органами (окучники, долота, стрельчатые лапы). Уход за картофелем на тяжелых почвах: В случае применения фрезерных культиваторов уход за картофелем сводится к одной процедуре - формирование гребней высотой 20 — 25 см. Основной уход за картофелем включает следующие виды работ: — борьба с сорняками; — окучивание с целью формирования гребней; — поддержание почвы в рыхлом состоянии. При выпадении обильных осадков междурядья рыхлят ярусными окучниками, чтобы не было трещин.

Ключевые слова	Информативность
ярусными	8.33
насыпается	8.33
стрельчатые	3.45
рыхлители	2.38
рыхлят	2.04
долота	2.00
бороны	1.92
окучниками	1.54
боронования	1.28
обрабатываются	1.20

Рисунок 3. – Результат реферирования
Figure 3. – The result of summarizing

Индексирование веб-страниц

Целью индексирования веб-страниц является приписывание им совокупностей ключевых слов с их весами. Вес – это информативность слова. В поисковый образ веб-страницы включаются те слова, информативность которых не меньше некоторого порогового значения.

Алгоритм индексирования функционирует в два шага. На первом шаге вычисляется информативность каждого слова на веб-странице. Если информативность не меньше некоторого порогового значения, то слово является ключевым. На втором шаге из пар <ключевое слово, информативность ключевого слова> формируется поисковый образ веб-страницы.

Критерий выдачи

Под критерием выдачи понимается правило, по которому вычисляется степень соответствия запросу веб-страниц или текстовых документов, найденных в процессе информационного мониторинга Интернета и поиска в полном корпусе текстов. В данной системе интернет-мониторинга используется векторная модель поиска. Запрос пользователя и поисковый образ веб-страницы представляются в виде векторов n-мерного евклидова пространства, в роли координат которых выступают значения информативности соответствующих ключевых слов. В качестве критерия выдачи применяется косинус угла между векторами поискового предписания и поискового образа веб-страницы.

Документальный поиск

Информационный поиск в множестве веб-страниц реализуется в четыре этапа. На первом этапе формируется динамический корпус текстов, релевантный запросу пользователя. На втором этапе корректируется исходный запрос. Документы из динамического корпуса предъявляются пользователю, который исключает из него все непертинентные тексты. Полученное в результате множество считается уточненным динамическим корпусом текстов, на основе которого путем его индексирования формируется уточненное поисковое предписание. На третьем этапе проводится поиск по новому поисковому предписанию. На четвертом этапе результаты поиска ранжируются с использованием принятого в системе критерия выдачи.

Фактографический поиск

Поиск сводится к выделению информативных предложений в тексте найденной веб-страницы, релевантных построенному динамическому корпусу текстов. Процедура реализуется в три этапа. На первом этапе вычисляется информативность каждой словоформы текста. На втором этапе определяется информативность всех предложений. На третьем этапе находятся все предложения текста, информативность которых не меньше некоторого значения. Совокупность этих предложений является результатом фактографического поиска.

Заключение

Определяющую роль при планировании архитектуры системы интернет-мониторинга научно-технической информации и ее отдельных компонентов играет состав задач. К числу наиболее актуальных можно отнести следующие задачи информационного мониторинга:

– подборка веб-страниц по запрашиваемой тематике (информирование пользователей о новых публикациях в их предметных областях, информация для принятия решений, деловая и экономическая разведка, тенденции развития и состояния рынков товаров и услуг и т. п.);

– информационный мониторинг текстовых документов по запрашиваемой тематике (подборка информативных выдержек из веб-страниц, дайджест новостей, выявление тонально окрашенной информации, т. е. публикаций экстремистского содержания, «электронных сплетен», инсинуаций и т. п.);

– классификация и рубрикация найденной информации;

– поиск документов по рубрикам;

– контроль обновляемости сайтов Интернета;

– формирование отчетов по результатам мониторинга.

Список использованных источников

1. Ненахова, А. Мониторинг информации [Электронный ресурс] / А. Ненахова, П. Васильев // Директор информ. службы. – 2007. - № 1. - Режим доступа: https://www.osp.ru/cio/2007/01/3923816. - Дата доступа: 11.04.2023.
2. Базаров, Р. Агентура деловых спецслужб [Электронный ресурс] / Р. Базаров. - Режим доступа: http://ci-razvedka.ru/Agentura_Delovykh_Specsluzhb.html. - Дата доступа: 10.04.2023.
3. Тактаев, С. Поиск информации в компьютерных сетях: новые подходы [Электронный ресурс] / С. Тактаев // SearchEngines. - Режим доступа: http://www.searchengines.ru/articles/004603.html. - Дата доступа: 10.04.2023.
4. Буравкин, А. Г. Интернет-поиск альтернативных вариантов в процессе принятия решений / А. Г. Буравкин, С. Ф. Липницкий, Л. В. Степура // Развитие информатизации и государственной системы научно-технической информации (РИНТИ-2021) : докл. XX Междунар. науч.-техн.

конф., Минск, 18 нояб. 2021 г. / Объед. ин-т проблем информатики Нац. акад. наук Беларуси ; науч. ред.: А. В. Тузиков, Р. Б. Григянец, В. Н. Венгеров. – Минск, 2021. – С. 180–183.

References

1. Nenakhova A. Information monitoring. *Direktor informatsionnoi sluzhby* [Information Service Director], 2007, no. 1. Available at: https://www.osp.ru/cio/2007/01/3923816 (accessed 11.04.2023) (in Russian).
2. Bazarov R. *Business intelligence agency.* Available at: http://ci-razvedka.ru/Agentura_Delovykh_Specsluzhb.html (accessed 10.04.2023) (in Russian).
3. Taktaev S. *Information search in computer networks: new approaches.* Available at: http://www.searchengines.ru/articles/004603.html (accessed 10.04.2023) (in Russian).
4. Buravkin A., Lipnitsky S., Stepura L. Internet search for alternatives in the decision-making process. *Razvitie informatizatsii i gosudarstvennoi sistemy nauchno-tekhnicheskoi informatsii (RINTI-2021): doklady KhKh Mezhdunarodnoi nauchno-tekhnicheskoi konferentsii, Minsk, 18 noyabrya 2021 g.* [Development of Informatization and the State System of Scientific and Technical Information (RINTI-2021): reports of the XX International Scientific and Technical Conference, Minsk, November 18, 2021]. Minsk, 2021, pp. 180–183 (in Russian).