

ТЕХНИЧЕСКОЕ ОПИСАНИЕ И ХАРАКТЕРИСТИКИ СИСТЕМЫ АВТОМАТИЗАЦИИ ИНФОРМАЦИОННО-АНАЛИТИЧЕСКОЙ РАБОТЫ INFOSTREAM CORPORATE

1. НАЗНАЧЕНИЕ СИСТЕМЫ

1.1. Система InfoStream Corporate (далее - Система) предназначена для автоматического сбора информации из сети Интернет из определенных Заказчиком источников (веб-сайтов), ее структурирования, группирования по семантическим признакам, избирательного распределения информации, а также предоставления доступа к ней в режимах поиска и автоматического обобщения (аналитической обработки).

1.2. Система обеспечивает:

- возможность оперативного автоматического сбора информации с Интернет-ресурсов;
- интерактивную настройку на сканирование выбранных веб-сайтов для обеспечения сбора актуальной информации;
- преобразование потоков документов из Интернет в «информационные ленты» унифицированного текстового вида;
- формирование и ведение оперативных и ретроспективных баз данных;
- выявление (экстрагирование) из документов заданных понятий;
- интерактивное ведение запросов пользователей;
- эффективный одновременный доступ к базам данных многим пользователям в режимах поиска и обобщения информации;
- оперативную доставку пользователям релевантных документов;
- надежное хранение документов;
- возможность интерактивного поиска и анализа состояния потоков документов из Интернет;
- автоматическое обобщение информации, формирование «сюжетных цепочек», дайджестов, информационных портретов, ситуационных карт и т.д.

1.3. Система автоматизации информационно-аналитической работы InfoStream Corporate реализована на базе программно-технологического обеспечения системы интеграции и мониторинга ресурсов сети Интернет InfoStream и полнотекстовой информационно-поисковой системы InfoRes. Программно-технологическое обеспечение InfoStream дает возможность

осуществлять информационно-аналитическую поддержку пользователей при работе с информационными ресурсами сети Интернет, которые соответствуют сфере их интересов.

1.4. В результате внедрения системы автоматизации информационно-аналитической работы на базе программно-технологического обеспечения InfoStream и InfoRes реализуется:

- эффективный поиск информации из большого количества интернет-ресурсов;
- автоматическое обобщение больших и постоянно возрастающих объемов необходимой информации по интересующим пользователей проблематикам;
- автоматическая структуризация собираемой и обрабатываемой информации;
- эффективный доступ пользователей к информации по интересующим его вопросам как в отечественных, так и в зарубежных источниках;
- автоматическая фильтрация и существенное уменьшение информационного шума.

1.5. Система обеспечивает автоматизацию доступа к информационным потокам с веб-сайтов сети Интернет, чем способствует улучшению результатов работы пользователей за счет обеспечения таких показателей, как оперативность, полнота и точность поиска.

2. ОБЛАСТЬ И УСЛОВИЯ ПРИМЕНЕНИЯ

2.1. Система автоматизации информационно-аналитической работы необходима для использования в организациях, осуществляющих обработку и обобщение больших объемов информации из большого количества источников с целью поддержки подготовки и принятия решений.

2.2. Использование Системы пользователями осуществляется как в пакетном (оффлайн), так и в интерактивном (онлайн) режимах.

2.3. Эффективность Системы обеспечивается за счет реализации возможностей:

- современных технологий информационного поиска и глубинного анализа текстовых массивов большого объема;
- эффективного использования средств интеграции и мониторинга информационных ресурсов сети Интернет для поддержки информационно-аналитической деятельности;
- коллективной работы с базами данных текстовой информации и отдельными документами;
- эффективного персонализированного доступа к информационным материалам;
- современных элементов веб-технологий;
- средств защиты данных и информационной безопасности, обеспечения целостности информации и системного сопровождения.

2.4. Программное обеспечение Системы – это серверное решение, реализуемое на нескольких серверах. Минимальное количество серверов для обеспечения функционирования системы - два, допустимая спецификация которых должна быть не ниже, чем:

- 1-й сервер: Intel Pentium IV, RAM 2 Gb, HDD 2x73 Gb SCSI, 2xLan Ethernet 100 Мб;
- 2-й сервер: Intel Pentium IV, RAM 2 Gb, HDD 2x500 Gb SATA, Lan Ethernet 100 Мб.

В случае необходимости количество серверов может быть увеличено до четырех. Необходимое количество серверов для обеспечения высокоэффективной работы Системы в значительной мере зависит от объемов обрабатываемых информационных потоков и количества одновременно работающих пользователей и от требований по надежности функционирования.

Операционная система, в среде которой функционирует программное обеспечение системы автоматизации информационно-аналитической работы - ОС FreeBSD 6.3 и выше.

3. СОСТАВ И СТРУКТУРА СИСТЕМЫ

3.1. В состав системы автоматизации информационно-аналитической работы входят такие функциональные подсистемы:

- сбора и первичной обработки информации (Подсистема 1);
- избирательного распределения информации (Подсистема 2);
- онлайн-доступа к данным (Подсистема 3);
- аналитических средств (Подсистема 4);
- системного администрирования (Подсистема 5).

3.2. Каждая подсистема, в свою очередь, может быть представлена в виде совокупности отдельных модулей.

3.3. Структура Системы для двухсерверной конфигурации приведена на рис. 1.

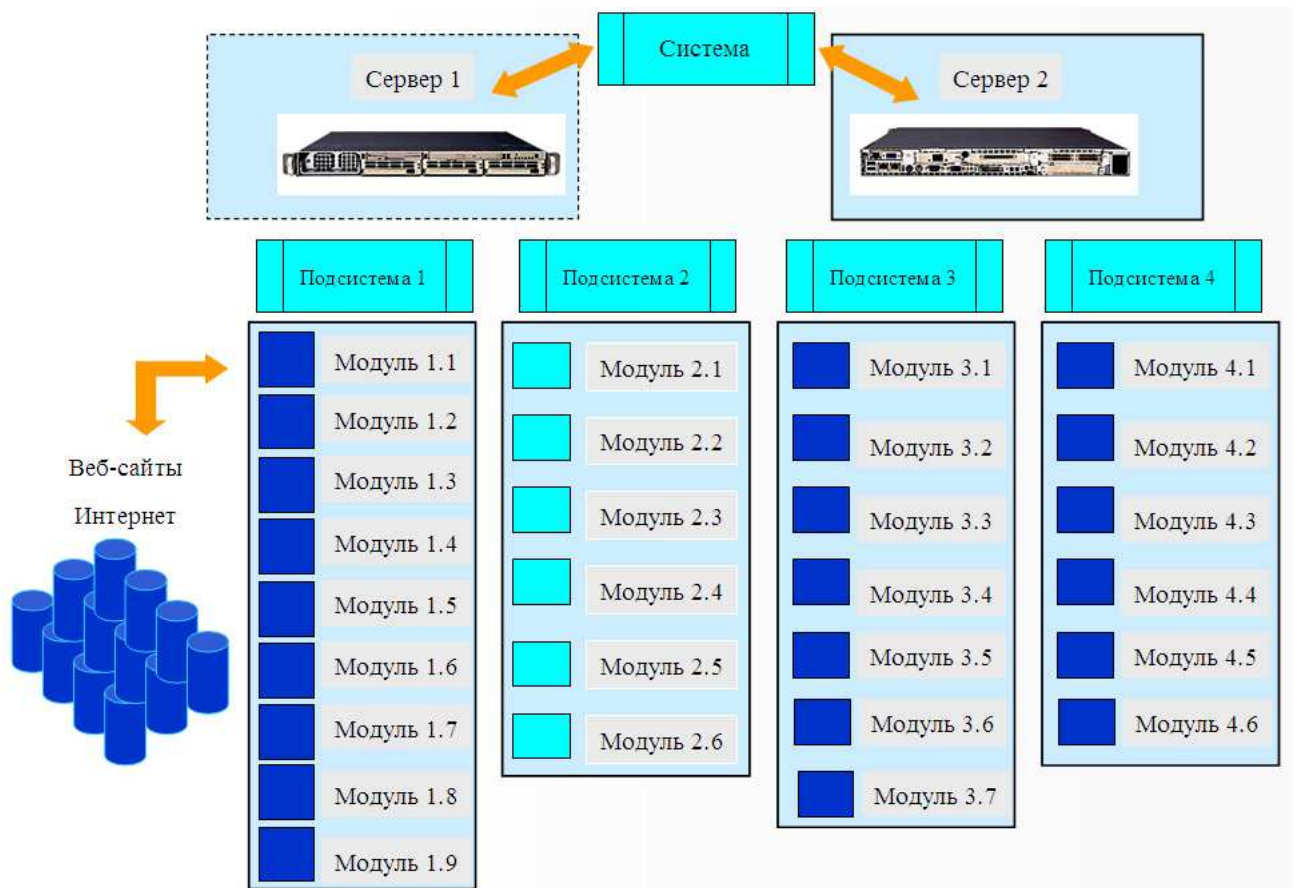


Рис. 1. Структурная схема Системы для двухсерверной конфигурации

3.4. Подсистемы 1 и 2 размещаются на первом сервере, на втором сервере размещаются подсистемы 3 и 4. Компоненты Подсистемы 5 присутствуют на двух серверах.

4. ОПИСАНИЕ ФУНКЦИОНИРОВАНИЯ СИСТЕМЫ

4.1. Подсистема сбора и первичной обработки информации

4.1.1. Подсистема сбора и первичной обработки информации принимает входных поток, преобразует его во внутренний системный формат информационно-поисковой системы InfoRes, выявляет признаки отдельных полей и проводит разметку текста, экстрагирует ключевые слова - наиболее весомые по лингвостатистическим критериям слова из сообщений, а также выполняет их нормализацию. Эта подсистема также выявляет содержательные дубликаты документов, которые поступают на вход системы.

4.1.2. Для сбора информации в Подсистеме 1 применяется модуль сканирования и первичной обработки информации с веб-сайтов (Модуль 1.3). Канонический вид сообщения, которое формируется этим модулем - заголовок, аннотация и тело сообщения. Все другие структурные части сообщения, если они есть (дата публикации, автор, первоисточник сообщения, принадлежность выпуску и т.д.), включаются в тело сообщения. Все сообщения, преобразованные в текстовый вид, образуют входной поток.

4.1.3. Для макроописания ресурсов, размещенных на веб-сайтах сети Интернет и создания соответствующих конфигураторов используются Модуль 1.1 и Модуль 1.2. Ресурсы, для которых составлены макроописания, и подключенные к системе, должны сканироваться и приводиться к стандартному текстовому виду с помощью Модуля 1.3 - сканирования и первичной обработки информации с веб-сайтов. Также с помощью этого модуля документам приписываются такие параметры, как URL, название источника, язык сообщения, страна, дата и время сканирования.

4.1.4. Модуль 1.6 предназначен для извлечения (экстрагирования) ключевых слов, наиболее весомых по лингвостатистическим критериям, при этом формируется словарь потока. Создание этого словаря связано с процедурой индексации сообщений. Словарь потока используется в дальнейшем для присвоения отдельным документам признака “дубликат”, а также в поисковом механизме и в подсистем аналитических средств.

4.1.5. Модуль 1.7 используется для выявления содержательного дублирования документов. Этот модуль обрабатывает входной поток, добавляя к сообщению признак «дубликат», если среди уже обработанных сообщений есть содержательные повторы текущего сообщения.

4.1.6. Модуль извлечения понятий из полнотекстовых документов (Модуль 1.4) предназначен для расширения сообщения такими полями: фамилии персон, географические данные, названия компаний, которые упоминались в документе. Модуль определения тональности сообщений (Модуль 1.5) базируется на байесовском подходе к определению

эмоциональной окраски документа. Модуль статической рубрикации (Модуль 1.8) обеспечивает классификацию документов в соответствии с типовыми запросами, которые соответствуют информационным потребностям пользователей. Предусмотрены также средства ведения базы данных запросов для такой рубрикации (Модуль 1.9).

4.2. Подсистема избирательного распределения информации

4.2.1. Подсистема избирательного распределения информации предназначена для отбора и распространения информации по предварительно установленным запросам пользователей с помощью электронной почты и сообщений SMS. Эта подсистема состоит из таких модулей: ведения базы данных запросов пользователей (Модуль 2.1); поиска и распределения данных (Модуль 2.2); ведения базы данных пользователей (Модуль 2.3); сбора и обработки статистики (Модуль 2.4); ведения почтовых ящиков пользователей (Модуль 2.5); информирования пользователей средствами SMS (Модуль 2.6).

4.3. Подсистема онлайн-доступа к данным

4.3.1. Подсистема онлайн-доступа к данным обеспечивает создание и ведение баз данных, а также обеспечение доступа к ним со стороны пользователей в режиме онлайн. После сканирования и предварительной обработки входной поток размещается в полнотекстовых базах данных. Поскольку сообщения поступают в полнотекстовые базы данных постоянно, то периодически выполняется процедура перемещения (ротации) части сообщений из оперативной базы данных в ретроспективные.

4.3.2. Модуль создания и ротации баз данных (Модуль 3.1) преобразует обработанный входной поток в полнотекстовые базы данных прямого доступа. Ротация баз данных связана с тем, что в оперативной базе данных хранятся документы за определенный промежуток времени.

4.3.3. С помощью Модуля 3.2 обеспечивается интерактивный поиск в режиме диалогового доступа к базам данных. В этом режиме пользователи могут осуществлять просмотр, поиск и отображение данных, а также получать доступ к оригиналам документов с веб-сайтов сети Интернет. Поисковый интерфейс позволяет вводить, редактировать, сохранять поисковые запросы, осуществляя таким образом конфигурацию информационного канала (определяемого сохраняемым запросом). Поисковый интерфейс предоставляет возможность определения критериев поиска, в том числе:

- поисковых терминов;
- логических операторов;

- признака учета морфологии;
- признака учета сообщений-дубликатов;
- временного периода поиска.

4.3.4. К критериям запроса относятся термины и параметры информационного портрета. Модуль построения информационного портрета (Модуль 3.3) обеспечивает выявление и визуализацию статистически весомых слов, рубрик, источников и других параметров, которые сопровождают результаты поиска по информационным запросам, и могут детализировать запросы.

4.3.5. С помощью Модуля 3.5 обеспечивается адаптация интерфейса пользователя и его персонализация. Пользователь имеет возможность сохранения выбранных им источников информации и отлаженных запросов, так называемых персональных информационных каналов. Эти источники и запросы отображаются на главном интерфейсе пользователя. Пользователь имеет возможность корректировки сохраненных им запросов.

4.3.6. Доступ к базе данных информационных источников со стороны пользователей поддерживается с помощью Модуля 3.6. С помощью этого модуля пользователь имеет возможность осуществить поиск требуемого источника по названию или по фрагменту URL, просмотреть каталог источников по видам (информационные агентства, сетевые СМИ, газеты, периодические издания, веб-сайты компаний и т.п.) и перейти к режиму поиска информации по выбранному источником.

4.3.7. Модуль автоматического классификатора-навигатора (Модуль 3.4) используется для построения иерархического каталога на основе ключевых слов, входящих в релевантные документы. Классификатор-навигатор может применяться также для уточнения первоначально введенных запросов

4.3.8. Подсистема 3 содержит специальный модуль импорта пользователями результатов поиска в формате RSS (Модуль 3.7), который предоставляет возможность доступа к информации с помощью современных программно-технологических средств агрегации интернет-контента.

4.4. Подсистема аналитических средств

4.4.1. Подсистема аналитических средств реализует аналитически-обобщающие элементы концепции контент-мониторинга, с помощью которых обеспечивается решение задач формирования тематических сюжетных цепочек, дайджестов, ситуационных карт и т.п.

4.4.2. Подсистема 4 содержит такие модули: формирования сюжетов (Модуль 4.1); формирования дайджестов (Модуль 4.2); построения таблиц взаимосвязей понятий (Модуль 4.3);

построения ситуационной карты (Модуль 4.4); выявления динамики понятий (Модуль 4.5); статистической обработки информации (Модуль 4.6).

4.4.3. Модуль 4.1 используется для формирования сюжетов и реализует автоматическое объединение тематически близких документов, а также их содержательное ранжирование. Этот модуль обеспечивает формирование сюжетных цепочек, которые отражают все весомые сообщения по заданной тематике. Средства этого модуля базируются на теории семантических сетей и технологических решениях в области автоматического выявления наиболее взаимозависимых документов. Основные факторы, которые влияют на ранжирование - это размеры сюжетных цепочек и оперативность сообщений, которые в них представлены. Размер сюжетной цепочки отражает общий интерес к конкретной теме, а оперативность - новизну сюжета. Также обеспечивается анализ заголовков документов, которые входят в сюжеты, в результате чего из всех заголовков выбираются наиболее адекватные для отображения. Построение перечня основных сюжетных цепочек существенно упрощает механизмы взаимодействия пользователя с подсистемой.

4.4.4. С помощью Модуля 4.2 на основе методов автоматического реферирования формируются дайджесты, которые обобщают массивы из большого числа документов. Дайджест в Системе рассматривается как аннотированный источник гиперссылок на документы, которые положены в его основу. Для построения дайджеста должны автоматически выбираться документы, в которых наиболее явным образом отражены тенденции всего входного потока. Для формирования дайджестов в данном модуле применяются статистические алгоритмы, основанные на частотном подходе.

4.4.5. Модуль 4.3 используется для построения таблиц взаимосвязей понятий. Этот модуль позволяет выявлять взаимосвязи таких понятий, как географические названия, персоналии, организации, которые присутствуют в текстовых документах - результатах поиска по первичному запросу. С помощью этого модуля реализуются процедуры построения таблиц взаимосвязи понятий, их перегруппировки и визуализации.

4.4.6. Модуль 4.4 реализует построение ситуационной карты, представляющей собой таблицы, в которых ячейками являются объекты предметной области, а колонки отражают отдельные характеристики этих объектов. По желанию пользователя, который может отмечать отдельные ячейки этой таблицы, рассчитываются и отображаются в виде таблиц информационные связи (количество вхождений в одни документы) между объектами.

4.4.7. Модуль 4.5 обеспечивает выявления динамики понятий и позволяет отслеживать появление сообщений по определенной пользователем проблематике за определенный период времени. Визуализация динамики понятий реализована в виде соответствующих диаграмм, в том числе с отображением тональности отдельных сообщений.

4.4.8. Модуль 4.6 обеспечивает статистическую обработку информации и является логическим расширением модуля 4.5. Для статистической обработки рядов, которые соответствуют интенсивности публикации по определенным понятиям и объектам, применяются методы статистического и корреляционного анализа.

5. ХАРАКТЕРИСТИКИ СИСТЕМЫ

5.1. Общие характеристики программного обеспечения

5.1.1. Программное обеспечение Системы реализует и поддерживает:

- архитектуру открытых систем;
- современные технологии сетей Internet/Intranet;
- ввод-вывод информации в интерактивном режиме;
- многовариантные процедуры поиска информации в базах данных;
- возможность модификации программного обеспечения;
- возможность дальнейшего развития или применения новых версий компонентов программного обеспечения;
- единые протоколы взаимодействия между компонентами программного обеспечения;
- максимальную унификацию для различных версий клиентского программного обеспечения;
- идеологическое единство компонентов программного обеспечения.

5.1.2. Системное программное обеспечение состоит из:

- операционной системы ОС FreeBSD;
- веб-сервера Apache;
- системы управления базами данных СУБД MySQL;
- интерпретатора языка программирования Perl.

5.1.3. Программное обеспечение Системы базируется системе интеграции и мониторинга интернет-ресурсов InfoStream и на отдельных модулях полнотекстовой информационно-поисковой системы InfoRes. Исключительное авторское право на эти программные продукты принадлежат ООО «Информационный центр «ЭЛВИСТИ» (свидетельства о регистрации авторских права на произведение 8381 и № 8379 от 22.09.2003, выданные Государственным департаментом интеллектуальной собственности МОН Украины и авторские договоры на передачу имущественных прав № 12/02/04-01 от 12.02.04, № 12/02/04-02 от 12.02.04).

5.2. Эксплуатационные характеристики

5.2.1. Эксплуатационные характеристики подразделяются на временные, качественные, объемные и количественные.

5.2.2. К временным характеристикам Системы относятся:

- периодичность сбора новых сообщений;
- скорость индексирования сообщений и размещения их в полнотекстовых базах данных;

- скорость обработки документов по пакету запросов в режиме избирательного распределения информации;
- время выполнения поискового запроса в режиме онлайн.

5.2.2.1. Периодичность сбора сообщений определяется периодом сканирования выбранных источников и должна быть не меньше номинального периода публикации на источниках новых сообщений. Этот показатель индивидуальный для каждого источника и может составлять от 15 минут до одного раза в сутки.

5.2.2.2. Автоматизированное создание оперативных баз данных должно осуществляться в режиме реального времени, время индексирования одного текстового документа объемом до 10 КБ не должно превышать 0,05 с.

5.2.2.3. Скорость обработки одного документа объемом до 10 КБ по одному запросу, входящему в пакет из 500 запросов объемом до 512 символов, в режиме избирательного распределения информации составляет 10000 документо-запросов за 1 с;

5.2.2.4. Время отклика информационно-поисковой системы на запрос в ретроспективной базе данных не превышает 30 с;

5.2.3. К качественным характеристикам Системы относятся:

- поддерживаемые языки представления документов;
- уровень полноты и точности результатов поиска;
- допустимое количество одновременно работающих с системой пользователей.

5.2.3.1. Лингвистическое обеспечение Системы в настоящее время позволяет работать с документами, представленными на русском, английском и украинском языках.

5.2.3.2. Система обеспечивает высокую точность выполнения поисковых запросов, выраженных на расширенном языке булевой логики. Полнота и точность удовлетворения информационных потребностей пользователей зависит от формулировки запросов к Системе. Для уточнения запроса может использоваться информационный портрет, соответствующий этому запросу, а также дополнительные аналитические средства, входящие в Систему.

5.2.3.3. Допустимое количество одновременно работающих с системой пользователей - не меньше чем 25.

5.2.4. К объемным и количественным характеристикам Системы относятся:

- количество сканируемых источников;
- количество информационных сообщений, которые поступают на вход Системы;
- количество запросов, которые должны обрабатываться Системой для минимально допустимой конфигурации технических средств;

- объем оперативной базы данных;
- объемы ретроспективных баз данных.

5.2.4.1. Количество сканируемых источников для минимально допустимой конфигурации технических средств составляет 5000.

5.2.4.2. Для канала связи с пропускной способностью 256 Кбит/с и сервера на платформе Intel Pentium IV допустимое количество входных сообщений может составлять не менее чем 20000 в сутки.

5.2.4.3. Допустимое количество запросов пользователей в режиме онлайн, которые должны обрабатываться системой для минимально допустимой конфигурации технических средств должны составлять не менее чем 250.

5.2.4.4. Объемы оперативных и ретроспективных баз данных не ограничиваются логически. Для минимально допустимой конфигурации технических средства объем баз данных составляет 50 млн. документов.