Rumyantsev Konstantin Evgenyevich – Southern Federal University; e-mail: ek.lozovskaya@yandex.ru; 2, Chekhov street, GSP-17A, Taganrog, 347928, Russia; phone: +78634371902; the department of information security of telecommunication systems; head of department; dr. of eng. sc.; professor.

Balabaev Sergey Leonidovich – the department of information security of telecommunication systems; cand. of eng. sc.; associate professor.

Lozovskaya Ekaterina Gennad'evna – the department of information security of telecommunication systems; postgraduate student.

УДК 004.67

В.В. Хашковский, А.Н. Шкурко

СОВРЕМЕННЫЕ ПОДХОДЫ В ОРГАНИЗАЦИИ СИСТЕМ ОБРАБОТКИ БОЛЬШИХ ОБЪЕМОВ ДАННЫХ

Обсуждаются современные подходы к организации систем обработки больших объемов данных на примере интегрированных систем известных производителей, и модульных решений независимых поставщиков. Основное внимание уделено методам получения и источникам информации для систем обработки данных. Приведены основные методы и источники получения исходной информации и дана их краткая характеристика. Рассмотрены основные этапы обработки информации в системах датамайнинга, начиная от непосредственно получения информации до формирования заключительного вывода по результатам анализа. Для основных этапов обработки приведены примеры существующих программных систем, реализующих необходимый функционал. Рассмотрены также некоторые подходы к определению характеристик документов и приведены примеры программных систем, реализующих эти подходы. Для исследуемых документов приведены основные параметры документов, на основании которых проводится анализ. В заключение делается вывод о состоянии рынка систем бизнес-анализа в России и перспектив их адаптации и внедрения.

Большие данные; датамайнинг; текстмайнинг; машинное обучение; классификация.

V.V. Khashkovsky, A.N. Shkurko

MODERN APPROACHES IN BIG DATA SYSTEMS

This paper discusses the current approaches to the organization of processing large amounts of data on an example of the integrated systems of leading manufacturers of modular and ISV solutions. The main attention is paid to the methods for the preparation and sources of information for data processing systems. The basic methods and sources of background information and a brief description of them is given. The main stages of information processing systems, data mining, ranging from direct information before forming a final conclusion on the results of the analysis. For the main processing steps are presented examples of existing software systems that implement the required functionality. We also consider some of the approaches to the characterization of documents and examples of software systems that implement these approaches. To study the documents shows the main parameters of documents on which the analysis is conducted. In conclusion we consider about the state of the market business intelligence systems in Russia and the prospects for their adaptation and implementation.

Big data; data mining; text mining; machine learning; classification.

Введение. Длительный период развития информационных технологий к настоящему времени привел к ситуации, которую можно охарактеризовать наличием большого объема накопленных в электронном виде данных, в том числе слабоструктурированных и разнотипных (аудио, видео, текст, базы данных). При этом лишь незначительная часть этих данных обладает определенной метаинформацией – в основном это относится к хорошо структурированным данным в среде баз данных.

Подавляющий объем данных либо обладает слабой метаинформацией, либо вовсе метаданные отсутствуют. Эта характеристика наилучшим образом подходит к данным, опубликованным в сети Интернет, что, с другой стороны, никоим образом не влияет на потенциальную их полезность. При этом сам по себе значительный накопленный объем информации позволяет ставить достаточно абстрактные задачи по анализу данных, которые с успехом могут быть решены человеком-экспертом, в помощь которому существуют и эффективно применяются системы анализа численных данных. Методы, которые при этом используются, в основном опираются на хорошо апробированный математический аппарат. Такие подходы группируются в основном в терминах Data Mining, Business Intelligence, Бизнес-аналитика и т.п. Рынок бизнес-аналитики превосходит по темпам роста весь рынок программного обеспечения в целом, который, по данным IDC, увеличился в 2012 г. на 3 %. Среди отдельных сегментов рынка бизнес-аналитики в 2012 г. наибольший прирост показал сегмент платформ хранения данных (10,8 % в годовом выражении). Сегмент непосредственно BI-систем и аналитических инструментов, а также сегмент управления эффективностью и аналитических приложений увеличились на 7,7 % каждый. На шесть вендоров при этом приходилось 64 % рынка в денежном выражении в 2012 г.: среди них такие компании, как Oracle, SAP, IBM, Microsoft, SAS, Teradata [1].

Однако это касается только численных данных. Неструктурированная текстовая информация пока что требует для выполнения анализа именно человека-аналитика, и значительный объем такой информации ставит вопрос о возможностях и подходах к автоматизации такого анализа и определении границ применения методов автоматизированной обработки. С другой стороны, большая часть информации в сети Интернет является именно текстовой и слабоструктурированной, что в свою очередь затрудняет применение классических BI-систем для обработки и анализа информации. При этом наиболее острой проблемой при внедрении подобных систем является задача отбора и предварительной обработки документов для последующего анализа.

Источники данных. В современной литературе [2] предлагается достаточно простая схема датамайнинг-систем, которая при этом охватывает все необходимые этапы обработки данных и позволяет строить системы высокой сложности.

Любая обработка начинается с источника данных. Как правило, системы датамайнинга строятся вокруг уже имеющихся данных, а потому в качестве источника данных рассматриваются существующие базы данных. Также отмечается [2], что в таком случае источник данных может рассматриваться в более широком смысле и представлять собой любой модуль, предоставляющий данные в необходимом для анализа количестве.

Данные, получаемые из источника данных, преобразуются в форматы, подходящие для дальнейшей обработки, и складируются в структурированное хранилище данных (data warehouse). Осуществление подробного первичного преобразования структуры данных позволяет повысить эффективность дальнейшего применения методов датамайнинга.

На следующем этапе происходит применение методов датамайнинга для извлечения необходимой информации, которая может быть также сохранена в структурированное хранилище. Характер извлекаемой информации напрямую зависит от задачи, которая решается системой.

В дальнейшем извлеченная информация передается в модули оценки и представления данных. В ходе работы этих модулей может быть оценена пригодность выделенных сведений (причем как автоматически, так и при помощи аналитика). Данные, полученные на выходе данной стадии, можно считать полезным знанием, которое было «добыто» системой.

При этом, как правило, описание источников данных в литературе не детализируется, хотя это, очевидно, может накладывать определенные ограничения на последующие этапы обработки данных. Таким образом, целесообразно рассматривать следующие способы и источники получения исходных данных более детально:

- «прямое сканирование web-страниц» – такой источник данных предполагает использование модулей периодического мониторинга web-узлов по списку, который может составляться как вручную аналитиком, так и пополняться в автоматическом режиме системой;
- «данные социальных сетей» – здесь предполагается осуществление сканирования социальных сетей (либо данных, предоставляемых поставщиками контента социальных сетей) на предмет поиска необходимых сведений;
- «запросы на поисковые системы» – этот источник предполагает использование существующих поисковых машин для мониторинга предметной области по ключевым словам;
- «данные агрегаторов» – использование этого источника данных предполагает мониторинг тематических новостных ресурсов с дальнейшим сканированием тех узлов, на которые ссылается агрегатор.

Следует отметить, что использование широкого диапазона источников данных позволяет осуществлять первичную оценку найденных материалов, а также варьировать начальный уровень коэффициента доверия [3] к найденным данным.

Данные, полученные от источников, сохраняются в централизованном хранилище «сырых» данных, в которое сохраняются ссылочные связи между документами, а также исключаются повторы больших блоков данных.

Сохраненные в хранилище «сырые» данные используются в дальнейшем модулями экстракции документов для выделения значимой части тела документа (далее контента) и отделения его от незначимых данных (удаление форматирования, лишних заголовков, меню сайтов и т.п.), а также предварительного технического анализа документов (например, выявления используемых терминов, статистических характеристик, общего настроения текста и т.п.). Выделенные таким образом характеристики сохраняются в соответствующие хранилища для дальнейшей обработки.

Прямое сканирование web-страниц. Прямое сканирование web-страниц является наиболее простым и наглядным методом получения данных интернет-ресурсов. Эта подход очень популярен, а следовательно, существует большое количество готовых систем. Большинство из них обладают схожими функциями: мониторинг ресурсов, сбор данных, преобразование данных во внутренние структуры и их простейший анализ.

Такие системы на этапе конфигурирования требуют список сайтов, в которых будет осуществляться поиск. Как правило, пользователем осуществляется настройка шаблонов выборки или правил для каждого ресурса, по которым будет производиться поиск контента. Некоторые решения позволяют производить фильтрацию данных по ключевым словам. Примерами подобных систем являются Web Data Extractor [4], WebSunDew [5], Mozenda [6], ScreenScraper [7], Kapow Katalyst [8].

Данные социальных сетей. Социальные сети в настоящее время имеют огромное значение как средство быстрого распространения информации. Отличительной особенностью контента, который публикуется в социальных сетях, является его относительная упорядоченность, поскольку, как правило, различные записи в них имеют схожую структуру. Также нельзя не отметить, что контент соци-

альных сетей персонифицирован, что делает возможным более простое отслеживание источников той или иной информации. Кроме того, социальные сети являются средой, где информация появляется и обновляется быстрее всего, в отличие от тех же поисковых машин, индексирующих ресурсы глобальной сети по расписанию.

Получение информации из социальных сетей возможно различными способами:

- прямым сканированием страниц (наиболее универсальный способ получения контента, который тем не менее имеет существенные недостатки, связанные с необходимостью адаптации сканеров под возможные изменения пользовательского интерфейса сети, а также связанные с созданием дополнительной нагрузки на серверы сети, которая может быть расценена как DoS-атака и заблокирована);
- средствами программных интерфейсов социальных сетей (большинство социальных сетей предлагают программные интерфейсы к своим сервисам, которые можно использовать для получения информации, но при этом зачастую для этих интерфейсов имеются ограничения по количеству обрабатываемых запросов, что может сделать использование данного метода получения информации затруднительным);
- получением информации через агрегаторов данных социальных сетей (подобные агрегаторы являются посредниками в получении данных социальной сети. Они, как правило, сохраняют данные из социальных сетей на регулярной основе в свои хранилища и предоставляют эту информацию клиентам. Очевидными проблемами в данном случае могут быть задержки в появлении контента у агрегатора, а также неполные данные).

Примерами сервисов-агрегаторов данных социальных сетей являются следующие:

1) GNIP [19] – один из лидеров рынка по данному направлению. Специализируются на предоставлении данных с наиболее популярных социальных сетей в унифицированном формате (более 20 штук). Предоставляют как данные за какой-то период (например, заявляют возможность предоставить полный архив всех сообщений в Twitter), так и поток в реальном времени. Кроме того, есть возможность фильтрации выдаваемых данных по ключевым словам, языку, георасположению и т.д. Стоимость решения зависит от потребностей заказчика и определяется после консультаций и пробной эксплуатации. (Стартовый порог – 500 $).

2) Spinn3r [20] – поставщик данных с социальных сетей и новостных сайтов в унифицированном формате в реальном времени. Стоимость решения зависит от потребностей заказчика и определяется после консультаций.

3) OracleSocialCloud [21] – работают с данными из социальных сетей с точки зрения бренда какой-либо компании. Мониторинг настроений пользователей бренда, отслеживание состояния бренда, выявление тем, трендов и активных обсуждений.

4) DataSift [22] – поставщик данных из социальных сетей (~19 штук) в универсальном формате. Предоставляют как архив за 3+ года, так и поток данных в реальном времени. Также фильтруют и анализируют: настроения, высокоуровневое определение темы, выявление сущностей (продукты, компании, люди, места), тренды. Предоставляют решения как на уровне предприятия, так и индивидуальным пользователям.

Учитывая вышесказанное, можно сделать вывод, что наилучшие результаты в смысле максимального полного просмотра информации из социальных сетей может дать сочетание всех способов получения информации. Так, например, услуги агрегаторов могут быть использованы для получения исторических данных, в то время как новые записи могут быть получены с использованием программного интерфейса или сканирования сайта сети.

Запросы на поисковые системы. Существующие поисковые системы (такие, как Google, Bing, Yandex и др.) обладают достаточными вычислительными мощностями для просмотра всех материалов, опубликованных в сети Интернет. Роботы этих систем регулярно просматривают серверы сети и соответствующим образом обновляют поисковые индексы, что делает возможным быстрый поиск информации по заданной теме. Тем не менее их использование как единственного источника данных для решения поставленной задачи не совсем оправдано по следующим причинам:

- наличие задержки между обновлением информации на сайте источника и в поисковом индексе, несмотря на то, что современные поисковые системы стремятся к минимизации этого показателя. При этом сайты с меньшим количеством посетителей могут сканироваться реже, обладая при этом более точной и оперативной информацией;
- закрытый алгоритм определения релевантности найденных результатов: отсутствие деталей о работе алгоритма поиска и ранжирования найденных материалов не дает возможность ответить на вопрос, найдены ли все актуальные материалы по заданным ключевым словам;
- фильтрация результатов поиска: некоторые результаты могут быть исключены из поисковой выдачи по различным причинам (нарушение соглашений, публикация запрещенных с точки зрения поисковых систем материалов и т.п.).

Учитывая обозначенные выше проблемы, можно заключить, что поисковые системы могут быть использованы только как дополнительный источник получения информации.

Данные агрегаторов. В настоящее время все большую популярность приобретают ресурсы, агрегирующие контент по определенной тематике. Это могут быть новостные сайты [9], тематические сообщества, например, Habrahabr [10] и т.п. Целью подобных ресурсов является облегчение поиска интересующей пользователей информации. Использование данных сайтов в качестве источника данных позволяет получать материалы, которые прошли предварительную редакцию, что может существенно повысить их качество. С другой стороны, распространение подобных ресурсов приводит к дублированию статей в сети [9], что может существенно осложнить анализ. Следует также отметить, что выбор подобных сайтов осуществляется в зависимости от предметной области аналитиком, а работа с подобными ресурсами, как правило, происходит при помощи прямого сканирования веб-контента.

Тем не менее подобные источники могут быть использованы наряду с другими для усиления позиций поиска информации.

Определение характеристик документов. Очевидно, что текст в его исходной форме малопригоден для автоматического анализа. Это связано со многими факторами, например, наличием различных словоформ, синонимов, слов-паразитов и т.п. Поэтому текст, полученный из различных источников, должен проходить определенные этапы подготовки для последующего анализа. Определяются слова, предложения, части речи. Этот процесс носит название POS tagging [14]. Вычисляются статистические характеристики документа: количество ссылок на документ, частота использования слов, их положение в документе.

Проводятся специфические преобразования текста, с целью облегчить дальнейший анализ: удаляются стоп-слова (шумовые слова), устраняются орфографические ошибки и опечатки, применяются стемминг или лемматизация.

Удаление стоп-слов состоит в исключении из текста слов, которые бесполезны для алгоритмов анализа, так как не несут смысловой нагрузки, полезной информации. Удаление уменьшает объем текстовой базы, улучшает производительность алгоритмов. Пример стоп-слов слов: “Так как”, “или”, “в”, “которая”, “определенные”, “так называемый”.

Стемминг – это преобразование текста, в процессе которого каждое слово в тексте заменяется на его основную форму. В итоге разная форма одного и того же слова будет восприниматься алгоритмами анализа одинаковым образом: “важная”, “важнейший” заменяется на “важн”, будет подразумевать понятие “важность” в обоих случаях. Это упрощает обработку текста в дальнейшем. Существует несколько вариантов реализации стемминга: поиск основы слова в таблице поиска и метод усечения окончаний.

Лемматизация также приводит возможные формы слова в одну, называемую нормальной формой. Например, для существительных нормальная форма слова – это слово в единственном числе, именительном падеже.

Возможна гибридная реализация, комбинация нескольких подходов, например, лемматизация по словарю, затем эвристика для слов, отсутствующих в словаре, – для получения приемлемой производительности и эффективности.

На рынке представлено достаточно большое количество реализаций стеммеров. Отметим некоторые из них:

ApacheOpenNLP [15] – содержит широкий набор средств по анализу языка: разбиение слова, предложения, определение частей речи, стеммеры.

Natural Language Toolkit [16] – открытый проект (на языке Python). В основном используется для обучения и исследований, прототипирования. Чаще всего его используют, чтобы “погрузиться” в тему, но, судя по отзывам, в нем есть ряд проблем, которые не подходят для продуктовой реализации.

KNIME Textprocessing [17] – дополнение к проекту по обработке данных KNIME, специализированное для обработки текста.

Стеммер Портера [18] – классическая реализация стеммера; реализован при помощи метода усечения окончаний. Поддерживает несколько языков, в том числе и русский.

Нужно отметить, что выбор вычисляемых характеристик документов, а также этапов преобразования текста зависит от методов, которые будут использованы для дальнейшей обработки данных. Следует соблюдать баланс в выборе количества рассчитываемых характеристик, так как при большом их количестве требуется большое количество вычислительных ресурсов для их расчета и последующего хранения. При этом недостающие характеристики могут быть рассчитаны из исходных документов, но при большом их количестве подобная операция также может потребовать существенных затрат.

Анализ документов. В отличие от оценки релевантности, ранжирование определяет, насколько документ, принадлежащий заданной тематике, важен для показа в первую очередь. Результатом ранжирования является отсортированный список документов, где позиция документа в списке определяет ранг, степень полезности документа. Современные системы оперируют тысячами различных параметров для определения степени пригодности документа заданной тематике, однако одновременно для документов определенного типа определяются сотни параметров [11].

Параметры документа могут быть следующих типов:

- статистические свойства документов (например, частота употребления слов соответствующей тематики);
- семантические свойства (например, положительные или отрицательные отзывы о товаре);

- свойства самого ресурса, на котором расположен документ (частота упоминания ресурса в Интернете, дата создания документа, количество ссылок на документ в Интернете).

Для документов разного характера (сообщение в социальной сети Twitter и научная статья) могут применяться разные методы определения ранга, что приводит к появлению специальных групп методов ранжирования, агентов. Каждый агент работает только по своей группе документов, выдавая ранг каждого документа в унифицированном виде, что дает возможность сравнить ранг с рангами документов другого типа.

Говоря о вычислении ранга, часто упоминают функцию ранжирования, т.е. некоторую формулу, вычисляющую ранг по фиксированному количеству числовых характеристик документа. Например, Okapi BM25 [13] – функция ранжирования, используемая поисковыми системами для упорядочивания документов по их релевантности данному поисковому запросу. Однако вычисление такой функции само по себе не дает эффективного результата, поэтому такие функции применяются как один из агентов по определению ранга.

Основной проблемой при выборе метода и его реализации для решения данной задачи является тот факт, что компании-лидеры в этой области держат свои разработки в секрете. Это приводит к необходимости глубокого исследования различных методов, сравнению их между собой, с целью выбора наиболее оптимального варианта решения задачи.

Выводы. В настоящий момент на мировом рынке решений интеллектуального анализа данных доминируют BI-системы известных компаний, в то время как на российском по-прежнему большую часть занимают собственные разработки. Наблюдение за отечественным рынком ИТ-решений показывает, что его динамика практически всегда копирует динамику мирового рынка, но со средним временем задержки год-два.

Что касается используемых технологий, здесь российский рынок как раз стоит на пороге начала внедрения BI-систем третьей волны, способных искать скрытую информацию, строить предсказательную аналитику и проводить перекрестный анализ информации из совершенно несовместимых на первый взгляд источников данных.

Другими словами, пользователю будет предоставлена возможность выбора сценария развития ситуации, исходя из которого система сама проведет анализ накопленной информации, построит прогноз изменения ключевых показателей и предложит оптимальные варианты действий, которые бы привели к лучшему результату. То есть в итоге BI-система избавит пользователя от необходимости выполнения длительной рутинной работы по поиску причинно-следственных связей при анализе данных, передаче результатов работы одной системы в другую, контроле корректности загруженной информации и т.д. Эти задачи будут выполняться автоматически – от пользователя потребуется всего лишь в начале работы выбрать сценарий развития той или иной задачи, а в конце – наиболее понравившийся ему оптимальный вариант решения этой задачи [23].

Адаптация и внедрение существующих систем BI крайне сложно осуществить силами конечных пользователей или отдела аналитики. Как правило, для этого требуется привлечение услуг интеграторов. Это делает бюджет внедрения BI-систем для решения задачи отбора и анализа документов в сети Интернет соизмеримым с бюджетом разработки новой системы. С другой стороны, богатый набор существующих инструментов для решения отдельных этапов обработки в рамках комплексной системы позволяет говорить о том, что комплексная аналитическая система может быть построена в рамках разумных сроков и бюджета.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Business Intelligence, BI (мировой рынок). ИТ-Директору, BI, Рынки, Рынки, программное обеспечение, 2014. [Электронный ресурс]. – Режим доступа: http://www.tadviser.ru/ index.php/%D0%A1%D1%82%D0%B0%D1%82%D1%8C%D1%8F:Business_Intelligence,_BI_ (%D0%BC%D0%B8%D1%80%D0%BE%D0%B2%D0%BE%D0%B9_%D1%80%D1%8B %D0%BD%D0%BE%D0%BA).
2. *Jiawei Han, Micheline Kamber.* Data Mining: Concepts and Techniques Second Edition / USA Elsevier Inc., 2006. – 743 p.
3. Коэффициент доверия в экспертных системах, 2014. [Электронный ресурс]. – Режим доступа: http://www.aiportal.ru/articles/expert-systems/confidence-factor.html.
4. Web Data Extractor – Extract Email, URL, Meta Tag, Phone, Fax from Websites, 2014. [Электронный ресурс]. – Режим доступа: http://www.webextractor.com.
5. Web Scraping, Web Extraction, WebSundew, 2014. [Электронный ресурс]. – Режим доступа: http://www.websundew.com/.
6. Data Extraction, Web Screen Scraping Tool, Mozenda Scraper, 2014. [Электронный ресурс]. – Режим доступа: https://www.mozenda.com/pricing.
7. Screen-scraper: Data extraction software and services, 2014. [Электронный ресурс]. – Режим доступа: http://www.screen-scraper.com/download/choose_version.php.
8. Kapow Katalyst: The Leading Application Integration Platform for connecting cloud, mobile, social and big data – Kapow Software, 2014. [Электронный ресурс]. – Режим доступа: http://www.kapowsoftware.com/products/kapow-katalyst/index.php.
9. *Гершензон Л.* Новостные агрегаторы и онлайн-СМИ: жизнь вместе, 2009 [Электронный ресурс]. – Режим доступа: http://download.yandex.ru/company/Yandex_News_11_2009.pdf.
10. О сайте / Хабрахабр, 2014. [Электронный ресурс]. – Режим доступа: http://habrahabr.ru/info/about.
11. *Гулин А., Карпович П., Расковалов Д., Сегалович И.* Оптимизация алгоритмов ранжирования методами машинного обучения, 2009 [Электронный ресурс]. – Режим доступа: http://romip.ru/romip2009/15_yandex.pdf.
12. Коллаборативная фильтрация – Википедия, 2014. [Электронный ресурс]. – Режим доступа: http://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D0%BB%D0%BB%D0%B0%D0% B1%D0%BE%D1%80%D0%B0%D1%82%D0%B8%D0%B2%D0%BD%D0%B0%D1%8F_ %D1%84%D0%B8%D0%BB%D1%8C%D1%82%D1%80%D0%B0%D1%86%D0%B8%D1 %8F.
13. Okapi BM25 – Википедия, 2013. [Электронный ресурс]. – Режим доступа: http://ru.wikipedia.org/wiki/Okapi_BM25.
14. Part-of-speech tagging – Wikipedia, the free encyclopedia, 2014. [Электронный ресурс]. – Режим доступа: http://en.wikipedia.org/wiki/Part-of-speech_tagging.
15. Apache OpenNLP – Welcome to Apache OpenNLP, 2010. [Электронный ресурс]. – Режим доступа: http://opennlp.apache.org/.
16. Natural Language Toolkit – NLTK 3.0 documentation, 2013. [Электронный ресурс]. – Режим доступа: http://www.nltk.org/.
17. KNIMEtech KNIME Text Processing, 2014. [Электронный ресурс]. – Режим доступа: http://tech.knime.org/knime-text-processing.
18. Russian stemming algorithm. [Электронный ресурс]. – Режим доступа: http://snowball.tartarus.org/algorithms/russian/stemmer.html.
19. The Source for Social Data – Gnip, 2014. [Электронный ресурс]. – Режим доступа: http://gnip.com/.
20. Spinn3r: RSS Content, News Feeds, News Content, News Crawler and Web Crawler APIs, 2014. [Электронный ресурс]. – Режим доступа: http://www.spinn3r.com/.
21. Oracle Social Cloud, Social Relationship Management (SRM) Solutions | Oracle, 2014. [Электронный ресурс]. – Режим доступа: http://www.oracle.com/us/solutions/social/ overview/index.html.
22. Data Sift Powering the Social Economy, 2014. [Электронный ресурс]. – Режим доступа: http://datasift.com/.
23. *Северов М.* Ключевые игроки рынка BI: круг сжимается, Аналитические системы, Информационные технологии, 2008. [Электронный ресурс]. – Режим доступа: http://www.iteam.ru/publications/it/section_92/article_3625/.

REFERENCES

1. Business Intelligence, (mirovoy rynok). IT-Direktoru, BI, Rynki, Rynki, programmnoe obespechenie, 2014 [Business Intelligence, (world market). The CIO, BI, Markets, Markets, software, 2014]. Available at: http://www.tadviser.ru/ index.php/%D0%A1%D1%82%D0%B0%D1%82%D1%8C%D1%8F:Business_Intelligence,_BI_(%D0%BC%D0%B8%D1%80%D0%BE%D0%B2%D0%BE%D0%B9_%D1%80%D1%8B%D0%BD%D0%BE%D0%BA).
2. *Jiawei Han, Micheline Kamber*. Data Mining: Concepts and Techniques Second Edition, USA Elsevier Inc., 2006, 743 p.
3. Koeffitsient doveriya v ekspertnykh sistemakh, 2014 [The factor of trust in expert systems, 2014]. Available at: http://www.aiportal.ru/articles/expert-systems/confidence-factor.html.
4. Web Data Extractor – Extract Email, URL, Meta Tag, Phone, Fax from Websites, 2014. Available at: http://www.webextractor.com.
5. Web Scraping, Web Extraction, WebSundew, 2014. Available at: http://www.websundew.com/.
6. Data Extraction, Web Screen Scraping Tool, Mozenda Scraper, 2014. Available at: https://www.mozenda.com/pricing.
7. Screen-scraper: Data extraction software and services, 2014. Available at: http://www.screen-scraper.com/download/choose_version.php.
8. Kapow Katalyst: The Leading Application Integration Platform for connecting cloud, mobile, social and big data – Kapow Software, 2014. Available at: http://www.kapowsoftware.com/ products/kapow-katalyst/index.php.
9. *Gershenzon L.* Novostnye agregatory i onlayn-SMI: zhizn' vmeste, 2009 [News aggregators and online media: life together, 2009]. Available at: http://download.yandex.ru/ company/Yandex_News_11_2009.pdf.
10. O sayte / Khabrakhabr, 2014 [About the website, Habrahabr, 2014]. Available at: http://habrahabr.ru/info/about.
11. *Gulin A., Karpovich P., Raskovalov D., Segalovich I.* Optimizatsiya algoritmov ranzhirovaniya metodami mashinnogo obucheniya, 2009 [Optimization algorithms of ranking methods in machine learning, 2009]. Available at: http://romip.ru/romip2009/15_yandex.pdf.
12. Kollaborativnaya fil'tratsiya – Vikipediya, 2014 [Collaborative filtering - Wikipedia, 2014]. Available at: http://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D0%BB%D0%BB%D0%B0% D0%B1%D0%BE%D1%80%D0%B0%D1%82%D0%B8%D0%B2%D0%BD%D0%B0%D1%8F_%D1%84%D0%B8%D0%BB%D1%8C%D1%82%D1%80%D0%B0%D1%86%D0%B8%D1%8F.
13. Okapi BM25 – Wikipedia, 2013 Available at: http://ru.wikipedia.org/wiki/Okapi_BM25.
14. Part-of-speech tagging – Wikipedia, the free encyclopedia, 2014. Available at: http://en.wikipedia.org/wiki/Part-of-speech_tagging.
15. Apache OpenNLP – Welcome to Apache OpenNLP, 2010. Available at: http://opennlp.apache.org/.
16. Natural Language Toolkit – NLTK 3.0 documentation, 2013. Available at: http://www.nltk.org/.
17. KNIMEtech KNIME Text Processing, 2014. Available at: http://tech.knime.org/knime-text-processing.
18. Russian stemming algorithm. Available at: http://snowball.tartarus.org/algorithms/ russian/stemmer.html.
19. The Source for Social Data – Gnip, 2014. Available at: http://gnip.com/.
20. Spinn3r: RSS Content, News Feeds, News Content, News Crawler and Web Crawler APIs, 2014. Available at: http://www.spinn3r.com/.
21. Oracle Social Cloud, Social Relationship Management (SRM) Solutions | Oracle, 2014. Available at: http://www.oracle.com/us/solutions/social/ overview/index.html.
22. Data Sift Powering the Social Economy, 2014. Available at: http://datasift.com/.
23. *Severov M.* Klyuchevye igroki rynka BI: krug szhimaetsya, Analiticheskie sistemy, Informatsionnye tekhnologii, 2008 [Key market players BI: circle shrinks, Analytical systems, Information technology, 2008]. Available at: http://www.iteam.ru/publications/it/ section_92/article_3625/.

Статью рекомендовал к опубликованию д.т.н., профессор В.П. Карелин.

Хашковский Валерий Валерьевич – Южный федеральный университет; e-mail: vkhashkovsky@sfedu.ru; 347928, г. Таганрог, пер. Некрасовский, 44; тел.: 88634371746; кафедра математического обеспечения и применения ЭВМ; к.т.н.; доцент.

Шкурко Алексей Николаевич – e-mail: anshkurko@sfedu.ru; кафедра математического обеспечения и применения ЭВМ; к.т.н.; доцент.

Khashkovsky Valery Valer'evich – Southern Federal University; e-mail: vkhashkovsky@sfedu.ru; 44, Nekrasovskiy, Taganrog, 347928, Russia; phone: +78634371746; the department of software engineering; cand. of eng. sc.; associate professor.

Shkurko Alexey Nikolaevich – e-mail: anshkurko@sfedu.ru; the department of software engineering; cand. of eng. sc.; associate professor.

УДК 004.4, 004.7

Е.А. Пакулова

РАСПРЕДЕЛЕНИЕ МУЛЬТИМЕДИА-ТРАФИКА В ГЕТЕРОГЕННОЙ СЕТИ

Целью работы являлась разработка алгоритма распределения трафика мультимедиа в гетерогенной сети. В основу предлагаемого алгоритма легла идея распределения отправленных бит по нескольким потокам с максимально возможной пропускной способностью, что позволяет агрегировать пропускную способность нескольких сетей доступа и снизить потери. Предложенное решение позволяет использовать несколько беспроводных технологий связи в условиях недостаточной пропускной способности. Кроме того, в статье проведен анализ существующих решений, выделены их недостатки. В заключении описан прототип системы, основанный на использовании WLAN (IEEE 802.11g) и LAN, и даны результаты экспериментального исследования алгоритма.

Распределение трафика; агрегация пропускной способности; снижение потерь.

E.A. Pakulova

MULTIMEDIA-TRAFFIC ALLOCATION OVER MULTIPLE PATHS IN HETEROGENEOUS NETWORK

The main purpose of a project was development of Sender-Side Path Scheduling (SSPS) algorithm. The main idea of proposed algorithm is sending bit rate allocation through several paths with maximum available bandwidth. Thus it allows to aggregate bandwidth of several access networks and reduces losses. The proposed algorithm can be used with several wireless technologies and under throughput restrictions. Also the survey of existing method was done. At the current stage of an investigation the prototype of a simple system for transmission multimedia traffic was build (on the basis of WLAN (IEEE 802.11g) and LAN).

Traffic allocation; bandwidth aggregation; losses reduction.

Introduction. The use of a wide range of sophisticated personal wireless devices is becoming commonplace in society. According to the paradigm "Always Best connected" [1] users of such devices demand to be able to use their full capabilities anywhere and anytime in everyday life. At the same time providers of wireless networks offer us various technologies for data transmission. These technologies differ in terms of services: QoS characteristics (throughput, packet loss rate and latency), pricing, coverage and etc. One of the most common service is transmission of high quality video content. According to the "Cisco Visual Networking Index: Forecast and Methodology" video transmis-