

Чередниченко О.Ю.

Кандидат технических наук,
Национальный технический университет
«Харьковский политехнический институт»,
Доцент кафедры программной инженерии и
информационных технологий управления,
Харьков, Украина
olha_cherednichenko@mail.ru

Янголенко О.В.

Национальный технический университет
«Харьковский политехнический институт»,
Ассистент кафедры программной инженерии и
информационных технологий управления,
Харьков, Украина
olga_ya26@mail.ru

Гонтарь Ю.Н.

Национальный технический университет
«Харьковский политехнический институт»,
Аспирант кафедры программной инженерии и
информационных технологий управления,
Харьков, Украина
sobotovi4ka@mail.ru

МОЗГОПОДОБНЫЕ СТРУКТУРЫ ДЛЯ СБОРА И АВТОМАТИЗИРОВАННОЙ ПЕРЕРАБОТКИ БИЗНЕС- ИНФОРМАЦИИ

Аннотация. В работе рассматриваются проблемы сбора и переработки больших объемов бизнес-информации на предприятии. Сформулирована задача систематизации бизнес-информации с целью повышения ее ценности для принятия решений. Определены основные составляющие процесса систематизации информации. Проанализированы алгоритмы поиска бизнес-информации в веб-пространстве, а также методы ее кластеризации. Предложены интеллектуальные модели поиска веб-источников информации, ее извлечения и кластеризации на основе метода компараторной идентификации.

Ключевые слова: бизнес-информация, систематизация, кластеризация, веб-источник, компараторная идентификация.

Формул:9; рис.: 7, табл.: 1, библи.: 8

JEL классификация: C8, M1

Olga Cherednichenko

PhD (Technical Science),
National Technical University
"Kharkiv Politechnic Institute",
Associate Professor at Department of
Software Engineering and Management
Information Technologies,
Kharkiv, Ukraine
olha_cherednichenko@mail.ru

Olha Yanholenko

National Technical University
"Kharkiv Politechnic Institute",
Assistant lecturer at Department of
Software Engineering and Management
Information Technologies,
Kharkiv, Ukraine
olga_ya26@mail.ru

Yulia Gontar

National Technical University
"Kharkiv Politechnic Institute",
PhD student at Department of
Software Engineering and Management
Information Technologies,
Kharkiv, Ukraine
sobotovi4ka@mail.ru

**BRAIN-LIKE STRUCTURES FOR COLLECTION AND
AUTOMATED PROCESSING OF BUSINESS INFORMATION**

Abstract. The given work considers the issues of collection and processing of big volumes of business information at the enterprise. The problem statement of systematization of business information is done in order to increase its value for decision-making. The main processes of information systematization are defined. The algorithms of business information search on the web are analyzed as well as clustering methods. The intelligent models of search of the web sources of information, its retrieval and clustering are suggested based on the method of comparator identification.

Keywords: business information, systematization, clustering, web source, comparator identification.

Formulas:9; fig.: 7, tabl.: 1, bibl.: 8

JEL Classification: C8, M1

Введение. Ключевым фактором успеха предприятия на сегодня является информация, содержащаяся в разнообразных документах. Своевременное использование различных источников информации предоставляет организациям возможность повысить собственную конкурентоспособность и быть частью глобальной бизнес среды.

В этой связи задачи поиска актуальных источников информации, ее сбора и обработки приобретают особую значимость. Руководство организаций заинтересовано в своевременном анализе данных, касающихся как финансовой, так и нефинансовой сторон основных бизнес-процессов. Решение задач сбора и переработки бизнес-информации с помощью современных информационных технологий обеспечивает процесс принятия решений актуальными, точными и объективными данными, что позволяет повысить эффективность управления организацией.

Анализ исследований и постановка задачи. Крупные и даже средние предприятия на сегодня, как правило, сталкиваются с огромными объемами данных, которые необходимо собирать, обрабатывать и хранить с целью использования в производственных процессах и в процессе принятия

управленческих решений [Гонтарь 2014]. Кроме того, все чаще приходится сталкиваться с данными в различных форматах (текстовом, аудио, видео), которые далеко не всегда подразумевают ее структурированный вид.

Вопрос создания электронных документов, их внедрение и обработки сегодня интересует многих исследователей [Гонтарь 2014, Поляков 2012, Соловьев 2014]. Большое внимание уделяется возможности хранения электронных документов и акцентируется внимание на системах электронного документооборота [Соловьев 2014].

Кроме того, проблема большого объема информации также связана с возрастающей скоростью изменения этой информации. То есть, данные в различных источниках все быстрее обновляются, и соответственно свежие данные должны собираться с все большей частотой и вовремя доставляться их «потребителям» в виде конкретных сотрудников и руководящих лиц.

Зачастую для решения всех этих проблем организациям требуются специальные ресурсы в виде автоматизированных систем и физических средств обработки и хранения больших объемов данных, что вызывает дополнительные расходы [Соловьев 2014]. Тем не менее, использование большего объема бизнес-информации позволяет обоснованно осуществлять процессы планирования и разработки стратегии развития предприятия, отыскивать скрытые закономерности и дает возможности для автоматизации бизнес-процессов.

Одним из ключевых вопросов в обработке информации с целью ее использования для принятия решений является качество данных. Ценность и качество информации для управления может быть оценена с использованием финансовых и нефинансовых критериев [Laneу 2011]. Финансовые критерии, например, предполагают оценку стоимости информации для реализации определенного бизнес-процесса или стоимость ее приобретения с точки зрения конкурентов. Нефинансовые характеристики касаются релевантности, актуальности данных, их полноты и точности, а также своевременности и доступности. Информация, содержащаяся в различных источниках, может дублироваться и быть неточной. Поэтому ее обработка требует дополнительных усилий, направленных на выявление таких фактов.

Целью статьи является разработка принципов поиска, извлечения и кластеризации бизнес-информации на основе интеллектуальных моделей как основы прикладной информационной технологии сбора и переработки бизнес-информации.

Проблемы, связанные с обработкой больших объемов бизнес-информации, и с обеспечением ее качества, приводят к постановке задачи систематизации управленческой информации (рис. 1). Под систематизацией информации подразумевается своего рода классификация всех документов организации по различным группам. Чаще всего вся документация предприятия распределяется в соответствии с номинальной, предметной, тематической, хронологической, авторской и архивной классификацией.

Номинальная систематизация представляет собой распределение документов по их типу; предметная систематизация – это распределение различных документов по принадлежности к какому-либо конкретному делу; тематическая – по общей тематике; хронологическая систематизация информации – распределение документов по дате их создания; авторская – по имени автора документа; архивная – по срокам хранения документации.

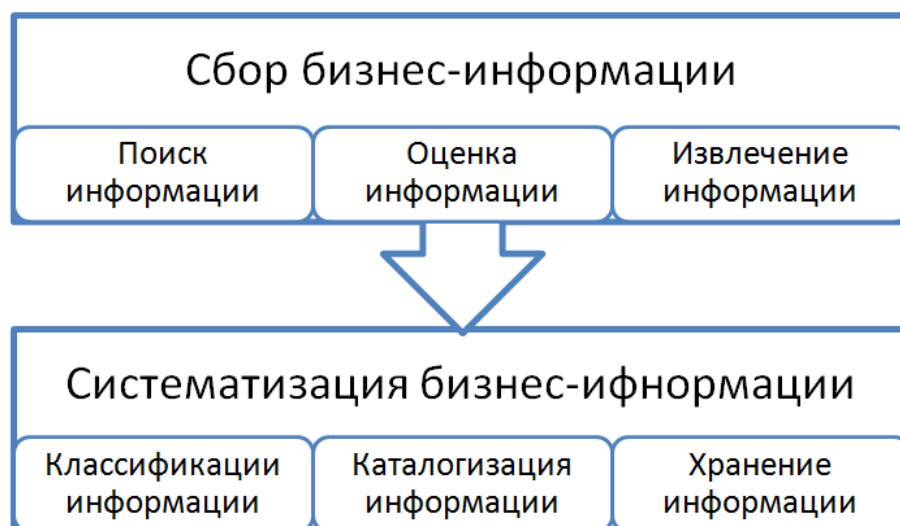


Рисунок 1 – Систематизация управленческой информации

Источник: авторская разработка

Целью систематизации документов является возможность их предоставления пользователям в краткие сроки. Для эффективного решения данной задачи необходимой является разработка моделей и информационной технологии сбора и переработки бизнес-информации в организации.

Результаты исследований.

Сбор бизнес-информации. В качестве источников данных бизнес-информации могут рассматриваться внутренние и внешние источники, которые характеризует внутреннюю и внешнюю среду предприятия (рис. 2).

Разнообразие источников данных свидетельствует о сложности процессов сбора и систематизации бизнес-информации. В данной работе рассматриваются внешние источники, а именно веб-ресурсы, содержащие электронные документы. Сбор данных из веб-пространства предполагает разработку моделей и алгоритмов поиска веб-страниц, содержащих определенную информацию о результатах деятельности предприятия.

Поиск и сбор веб-страниц осуществляется средствами веб-кроулинга, или обхода веб-пространства. Обход интернета (web crawling) – это процесс сбора веб-страниц в сети Интернет для их дальнейшего индексирования и поддержки функционирования поисковой системы [Davies 2006, Liu 2011]. Целью обхода является быстрый и эффективный сбор как можно большего количества полезных веб-страниц вместе со ссылками, которые их объединяют. Обход осуществляется поисковым роботом (веб-кроулером). Получая стартовый URL адрес, с которого начинается обход, поисковый робот загружает веб-страницу, добывает из нее все исходные ссылки и добавляет их в очередь для дальнейшего обхода. Этот процесс продолжается, пока очередь не окажется пустой.

Базовая схема работы поискового робота на практике расширяется функциональностью, связанной с оценкой релевантности веб-страницы, ее сохранением в базе данных и добычей других данных, кроме ссылок. Ключевым элементом функционирования поискового робота, который влияет на его эффективность, является стратегия обхода ссылок, которые хранятся в очереди [Liu 2011]. Чаще всего применяются два подхода: слепой поиск и эвристический подход.



Рисунок 2 – Источники данных бизнес-информации организации
Источник: авторская разработка

Во время слепого поиска при выборе следующего URL из очереди для загрузки не используется никакой критерий [Manning, Raghavan, Schütze 2009]. Ссылки для обхода выбираются в порядке их расположения в очереди. Наиболее распространенным алгоритмом слепого поиска является поиск в ширину (Breadth First Algorithm). Очередь на скачивание формируется в виде очереди типа FIFO (First In First Out) - «первым поступил - первым вышел». Ссылки обходятся в порядке, в котором они добавлялись в очередь. Недостаток этого алгоритма заключается в ограниченном объеме очереди, а также в необходимости сохранять потенциально нерелевантные ссылки.

Эвристический подход представлен методами, основанными на определенном критерии выбора следующей ссылки из очереди для обхода [Manning, Raghavan, Schütze 2009]. Эвристические алгоритмы классифицируются на те, которые требуют дополнительных знаний для определения критерия выбора ссылки из очереди и те, которые в них не нуждаются.

К алгоритмам, которые не требуют дополнительных знаний, принадлежит алгоритм первого наилучшего варианта (Best-First Algorithm), который базируется на том, что следующая ссылка выбирается из очереди на загрузку на основе определенной оценки или приоритета. То есть, каждый раз поисковый робот переходит на наиболее подходящую веб-страницу. Разновидностями данного алгоритма являются алгоритмы наивного первого наилучшего варианта (Naïve Best-First Algorithm), алгоритм ранжирования веб-страниц (PageRank Algorithm).

Сравнение алгоритмов обхода веб-страниц поисковым роботом приведено в таблице 1.

Результатом поиска веб-ресурсов бизнес-информации является неупорядоченный набор веб-страниц, которые могут содержать дублирующуюся, неточную информацию в неструктурированном виде. Поэтому следующими этапами сбора являются оценка найденной информации и ее извлечение.

Таблица 1 – Сравнение алгоритмов поиска источников информации в веб-пространстве

Алгоритм	Порядок обхода ссылок	Необходимые расчеты	Преимущества	Недостатки
Алгоритмы, не требующие дополнительных знаний				
Поиск в ширину	Ссылки обходятся в порядке, в котором они добавлялись в очередь (FIFO)	Не требует расчетов	Простота реализации	Ограниченный размер очереди, не учитывает тему поиска
Алгоритм PageRank	Выбирается ссылка с наибольшим PageRank коэффициентом	Расчет коэффициента PageRank	Учитывает популярность ссылок	Не подходит для решения задач тематического поиска
Алгоритм Batch-pagerank	Каждые K страниц упорядочиваются в соответствии со значением коэффициента PageRank	Расчет коэффициента PageRank для K страниц	Более быстрая реализации по сравнению с PageRank	Аппроксимация значения PageRank для выборки из K страниц не точная
Алгоритмы, требующие дополнительных знаний				
Алгоритм Fish-Search	Поиск в направлении страниц, релевантных поисковому запросу	Оценка релевантности в двоичной шкале	Динамический поиск с оценкой релевантности	Примитивная дискретная шкала оценивания релевантности
Алгоритм Shark-Search	Расширение Fish-Search нечеткими оценками от 0 до 1	Оценка релевантности учитывает унаследованную оценку	Дифференциация оценок, отсутствие параметра ширины поиска	Глубина поиска ограничивается нахождением нерелевантной страницы
Алгоритм тематического поиска	Ссылки обходятся в порядке FIFO с учетом оценки веб-страницы	Оценка перспективности страницы для дальнейшего поиска	Тематически-сфокусированный поиск	Расход времени на оценку страницы

Источник: авторская разработка

Нерелевантные данные должны быть отброшены еще на данных этапах, что приводит к необходимости осуществлять тематический

направленный поиск, основанный на оценке соответствия веб-страницы теме поиска.

Систематизация бизнес-информации. Огромные объемы найденной информации в сети Интернет часто приводят к тому, что количество объектов, которые удовлетворяют запросу пользователя, очень велико. Это усложняет процесс обзора результатов и выбора наиболее подходящих материалов (статей, обзоров, отчетов, др) из множества найденных. Однако в большинстве случаев огромные объемы информации можно сделать доступными для восприятия, если разбить источники (например, WEB-страницы) на тематические группы. Тогда, пользователь сразу может отбросить классы нерелевантных документов. Такой процесс группировки данных осуществляется с помощью кластеризации или классификации набора электронных документов.

В настоящее время существует множество методик, которые осуществляют группировки документов. Кластеризация (или кластерный анализ) – это задача разбиения множества объектов на группы – кластеры. Внутри каждой группы должны оказаться «похожие» объекты, а объекты разных группы должны быть как можно более различны [Everitt, Landau, Leese, Stahl, 2011]. Главное отличие кластеризации от классификации состоит в том, что перечень групп четко не задан и определяется в процессе работы алгоритма.

Применение кластерного анализа в общем виде сводится к следующим этапам:

1. Отбор выборки объектов для кластеризации.
2. Определение множества переменных, по которым будут оцениваться объекты в выборке, при необходимости – нормализация значений переменных.
3. Вычисление значений меры сходства между объектами.
4. Применение метода кластерного анализа для создания кластеров объектов.
5. Представление результатов анализа.

Наиболее распространенные алгоритмы кластеризации документов приведены на рис. 3. Каждый из алгоритмов требует вычисления меры сходства документов, которая определяется с помощью вектора характеристик для каждого объекта. В процессе нормализации все значения характеристик приводятся к некоторому диапазону.

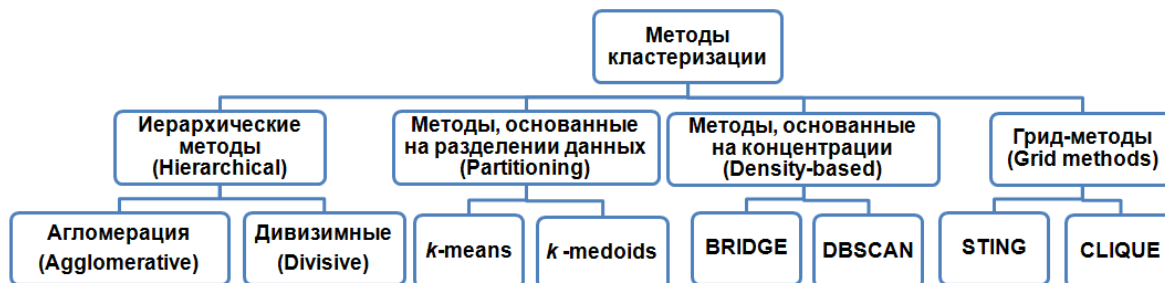


Рисунок 3 – Методы кластеризации бизнес-информации

Источник: авторская разработка

Наиболее распространенной мерой сходства можно считать геометрическое расстояние в многомерном пространстве:

$$\rho(x, x') = \sqrt{\sum_i^n (x_i - x'_i)^2}. \quad (1)$$

Квадрат евклидоваго расстояния

$$\rho(x, x') = \sum_i^n (x_i - x'_i)^2 \quad (2)$$

применяется для придания большего веса более отдаленным друг от друга объектам.

Расстояние городских кварталов (Манхэттенский расстояние) является средним разностей по координатам. Для этой меры влияние отдельных больших разностей уменьшается, так как они не возводятся в квадрат.

$$\rho(x, x') = \sum_i^n |x_i - x'_i|. \quad (3)$$

Расстояние Чебышева может оказаться полезным, когда нужно определить два объекта как «различные», если они различаются по одной координате:

$$\rho(x, x') = \max |x_i - x'_i|. \quad (4)$$

Степенное расстояние применяется в случае, когда необходимо увеличить или уменьшить вес, относящийся к размерности, для которой соответствующие объекты сильно отличаются:

$$\rho(x, x') = \sqrt[r]{\sum_i^n (x_i - x'_i)^p}, \quad (5)$$

где r и p – параметры, определяемые пользователем. Параметр p ответственен за постепенное взвешивание разностей по отдельным координатам, параметр r ответственен за прогрессивное взвешивание больших расстояний между объектами.

Применение традиционных мер сходства (1)-(5) для кластеризации электронных документов связано с представлением документа как объекта многомерного пространства, что приводит к необходимости поиска других интеллектуальных процедур оценивания схожести веб-источников.

Для автоматизации процессов поиска и кластеризации веб-ресурсов в данной работе предлагается моделировать эти процессы схожими с процессами человеческого мышления. Для этого предлагается использовать методы теории интеллекта [Бондаренко, Шабанов-Кушнаренко, 2011], в частности метод компараторной идентификации. Компаратор реализует предикат $K(y_1, y_2, \dots, y_m) = t$, что соответствует отношению K , в котором находятся входные сигналы y_1, y_2, \dots, y_m . При этом t – это двоичная реакция компаратора, $t \in \Sigma$, $\Sigma = \{1, 0\}$. К входам компаратора подключены своими выходами идентифицируемые информационные процессы $\Gamma_1, \Gamma_2, \dots, \Gamma_m$,

которые представляют механизмы восприятия входных физических сигналов x_1, x_2, \dots, x_m . Компаратор вместе с подключенными к нему информационными процессами называется идентифицируемым объектом (рис. 4) [Бондаренко, Шабанов-Кушнарченко, С.Ю., Шабанов-Кушнарченко, Ю.П., 2008].

Предикат объекта $P(x_1, x_2, \dots, x_m) = t$ выражается в виде $P(x_1, x_2, \dots, x_m) = K(r_1(x_1), r_2(x_2), \dots, r_m(x_m))$. Сигналы $y_1 = r_1(x_1)$, $y_2 = r_2(x_2)$, ..., $y_m = r_m(x_m)$ – это внутренние состояния объекта, недоступные для наблюдения.

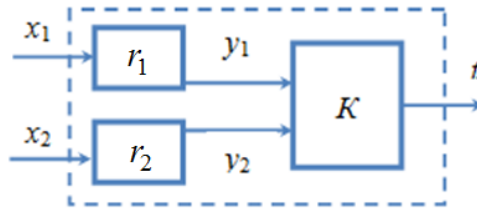


Рисунок 4 – Объект с двумя входными сигналами

Источник: авторская разработка

В данной работе предлагается осуществлять тематический поиск веб-источников бизнес-информации на основе веб-кроулера, который оценивает веб-страницу на основе компараторной модели (рис. 5). При этом компаратор сравнивает извлеченные из определенных элементов страницы слова со словами модели тематического поиска.

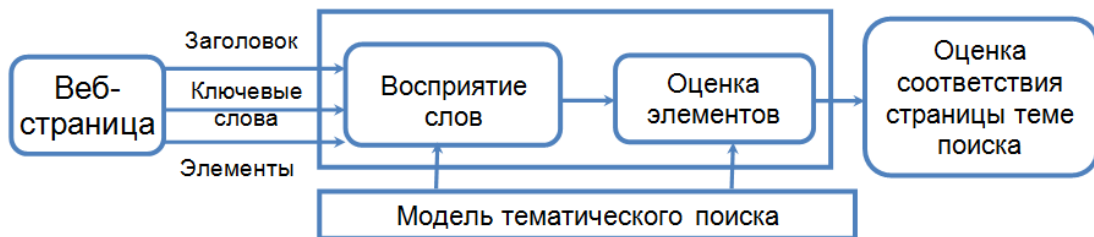


Рисунок 5 – Компаратор для поиска источников данных бизнес-информации

Источник: авторская разработка

Пусть E – множество структурных элементов веб-страницы, W – множество слов. Тогда $R_{\text{SEARCH}} \subseteq E \times W$ – бинарное отношение «используется для поиска». Обозначим $E_q \subseteq E$ – множество элементов страницы, выбранные для оценки и $W_q \subseteq W$ – множество слов, соответствующих теме поиска. Бинарное отношение $R_{\text{SEARCH}} = \{(e_{qi}, w_{qj}) \mid e_{qi} \in E_q, w_{qj} \in W_q\}$ задает пары «элемент-слово», для которых слова принадлежат множеству слов по теме поиска и элементы принадлежат множеству принятых для оценки элементов. Пусть $w_{pj} \in W_p$ – множество слов, извлеченных с веб-страницы. Тогда предикат, оценивающий бинарные пары «элемент-слово» задается как

$$P_w(e_{qi}, w_{pj}) = \begin{cases} 1, & (e_{qi}, w_{pj}) \in R_{\text{SEARCH}} \\ 0, & (e_{qi}, w_{pj}) \notin R_{\text{SEARCH}} \end{cases} \quad (6)$$

Предикат, определяющий наличие контрольных слов в определенном элементе имеет вид:

$$P_e(e_{qi}) = P_w(e_{qi}, w_{p1}) \vee P_w(e_{qi}, w_{p2}) \vee \dots \vee P_w(e_{qi}, w_{pn}). \quad (7)$$

Оценка веб-страницы объединяет оценки по каждому элементу:

$$P_q = P(e_{q1}) \vee P(e_{q2}) \vee \dots \vee P(e_{qs}). \quad (8)$$

Для извлечения данных из отобранных веб-страниц также предлагается использовать компараторную модель, которая позволяет для каждой страницы создать шаблон в соответствии с эталоном, которые содержит все необходимые структурные элементы (рис. 6). Эталонная структура созданных шаблонов используется при слиянии шаблонов разных страниц, содержащих дублирующуюся информацию.

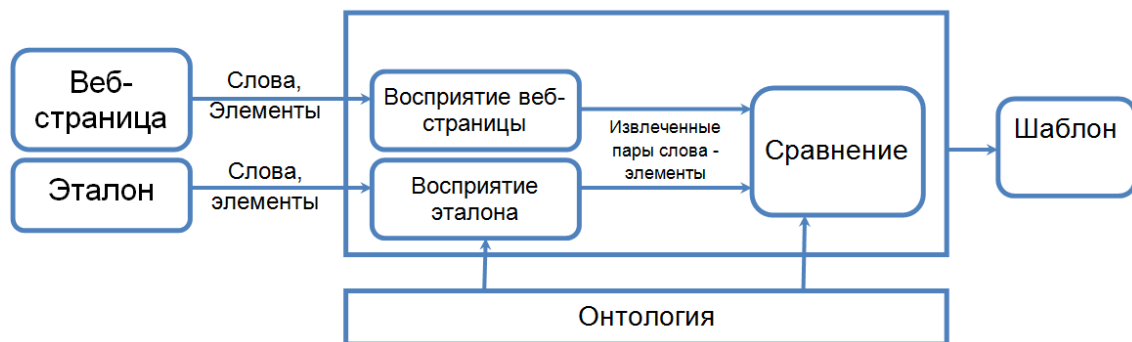


Рисунок 6 – Компаратор для извлечения данных из веб-страницы в шаблон
Источник: авторская разработка

Компараторная модель кластеризации сформированных шаблонов веб-источников предполагает их сравнение по нескольким дескрипторам, например, по наличию определенных слов в определенных структурных элементах шаблона (рис. 7).



Рисунок 7 – Компаратор для нахождения меры сходства веб-источников
Источник: авторская разработка

Мера сходства может быть определена как

$$m(d_1, d_2) = \sum a_i m_i(d_1, d_2), \quad (9)$$

где $m_i(d_1, d_2)$ определяет сходство документов d_1 и d_2 по i -му дескриптору; a_i – весовой коэффициент i -го дескриптора, $\sum a_i = 1$. Значение $m_i(d_1, d_2)$ определяется в соответствии с заданным правилом, например, $m_i(d_1, d_2) = 1$, если два документа имеют хотя бы одно общее слово хотя бы в одном структурном элементе.

Таким образом, компараторная модель позволяет представить процессы поиска бизнес-информации, ее извлечения и кластеризации подобными интеллектуальной деятельности человека.

Выводы. Рост значения бизнес-информации при принятии решений и построении программ развития предприятий становится все более ощутимым. При этом увеличивающаяся потребность в информации сопровождается ростом требований к ее качеству. Руководители предприятий нуждаются в свежей, полной и точной информации, доступной в форме удобной для анализа и использования.

Актуальность задач сбора и переработки бизнес-информации приводит к поиску новых решений, основывающихся на интеллектуальных информационных технологиях. В данной работе предложена реализация процессов поиска, извлечения и кластеризации бизнес-информации на основе интеллектуальных компараторных моделей. Метод компараторной идентификации позволяет осуществлять оценку электронных документов, их обработку и кластеризацию на подобие того, как это делает человек. Формальное представление моделей процессов систематизации бизнес-информации открывает возможности для автоматизации этих функций, чему будут посвящены дальнейшие исследования в данной области.

Литература

- Davies, J. (2006). *Semantic Web Technologies: Trends and Research in Ontology-based Systems*. Wiley.
- Everitt, B.S., Landau, S., Leese, M., Stahl, D. (2011). *Cluster analysis*. Wiley.
- Laney, D. (2011). Infonomics: The Economics of Information and Principles of Information Asset Management. *The Fifth MIT Information Quality Industry Symposium. Cambridge, 590-603*.
- Liu, B. (2011). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 2nd Edition*. Springer.
- Manning, C. D., Raghavan, P., Schütze, H. (2009). *An Introduction to Information Retrieval*. Cambridge, England: Cambridge University Press.
- Бондаренко, М. Ф., Шабанов-Кушнаренко, Ю.П. (2011). *Мозгоподобные структуры: Справочное пособие. Том первый*. Киев, Наукова думка.
- Бондаренко, М. Ф., Шабанов-Кушнаренко, С.Ю., Шабанов-Кушнаренко, Ю.П. (2008). Об общей теории компараторной идентификации. *Бионика интеллекта*, №2(69), 13-22.
- Гонтарь, Ю. Н. (2014). Принципы построения информационной технологии сбора и систематизации бизнес-информации. *Scientific Journal «ScienceRise»*, №5/2, 94-98.
- Поляков А. П. (2012). Информация и информационное обеспечение в системе контроллинга. *Культура народов Причерноморья*, № 234, 107-110.

Соловьев И.В. (2014). Каталогизация и индексирования информационных ресурсов. *Перспективы науки и образования*, №.4(10), 25-31.

References

- Bondarenko, M. F., & Shabanov-Kushnarenko, Yu.P. (2011). *Mozgopodobnyie strukturyi: Spravochnoe posobie. Tom pervyyi*. Kiev: Naukova dumka.
- Bondarenko, M. F., Shabanov-Kushnarenko, S.Yu., & Shabanov-Kushnarenko, Yu.P. (2008). Ob obschey teorii komparatornoy identifikatsii. *Bionika intellekta*, vol. 2(69), 13-22.
- Davies, J. (2006). *Semantic Web Technologies: Trends and Research in Ontology-based Systems*. Wiley.
- Everitt, B., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis*. Wiley.
- Gontar, Yu. N. (2014). Printsipyi postroeniya informatsionnoy tehnologii sbora i sistematzatsii biznes-ifnformatsii. *Scientific Journal «ScienceRise»*, vol. 5/2, 94-98.
- Laney, D. (2011). Infonomics: The Economics of Information and Principles of Information Asset Management. *The Fifth MIT Information Quality Industry Symposium. Cambridge*, 590-603.
- Liu, B. (2011). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 2nd Edition*. Springer.
- Manning, C., Raghavan, P., & Schütze, H. (2009). *An Introduction to Information Retrieval*. Cambridge, England: Cambridge University Press.
- Poljakov, A. P. (2012). Informacija i informacionnoe obespechenie v sisteme kontrollinga. *Kul'tura narodov Prichernomor'ja*, vol. 234, 107-110.
- Solov'ev, I. V. (2014). Katalogizacija i indeksirovanija informacionnyh resursov. *Perspektivy nauki i obrazovanija*, vol. 4(10), 25-31.

*Data przesłania artykułu do Redakcji: 20.06.2015
Data akceptacji artykułu przez Redakcję: 02.07.2015*