УДК 681.3.07

О.В Золотарёв[1]
Е.Б. Козеренко[2]
М.М. Шарнин[3]

ПРОВЕДЕНИЕ АНАЛИТИЧЕСКОЙ РАЗВЕДКИ НА ОСНОВЕ АНАЛИЗА НЕСТРУКТУРИРОВАННОЙ ИНФОРМАЦИИ ИЗ РАЗЛИЧНЫХ ИСТОЧНИКОВ, ВКЛЮЧАЯ ИНТЕРНЕТ И СРЕДСТВА МАССОВОЙ ИНФОРМАЦИИ

В статье рассматриваются вопросы проведения аналитической разведки на основе анализа текстов естественного языка. Описываются проблемы функционирования современного предприятия в среде Интернета. Анализируются тексты естественного языка в среде Интернета. Для проведения анализа текстов используется аппарат расширенных семантических сетей. В работе приведены примеры результатов проведения семантического анализа текстов.

Ключевые слова: *аналитическая разведка, бизнес-процесс, семантические сети, фрагменты знаний, объекты, тезаурус, большие объемы данных.*

O.V. Zolotarev
E.B. Kozerenko
M.M. Charnine

BUSINESS INTELLIGENCE PROCESSING ON THE BASE OF UNSTRUCTURED INFORMATION ANALYSIS FROM DIFFERENT SOURCES INCLUDING MASS MEDIA AND INTERNET

The article considers the issues of analytical intelligence based on the analysis of natural language texts. It also describes problems in the functioning of a modern enterprise in the Internet environment. Natural language texts are analyzed in the Internet environment. For the analysis of texts a mechanism of extended semantic networks is used. The paper presents the results of semantic analysis of texts.

Keywords: *business intelligence processing, business process, semantic networks, fragments of knowledge, objects, thesaurus, big data.*

Находясь в жесткой конкурентной борьбе, руководство предприятия вынуждено использовать различные аналитические и интеллектуальные методы для защиты своей репутации, охраны конфиденциальной информации, выявления разного рода утечек документов, чтобы минимизировать возможный ущерб от несанкционированной деятельности как своих работников, так и представителей конкурирующих фирм. Для этого проводятся аналитические исследования информации, получаемой из различных источников, включая Интернет. Отслеживаются различные значимые процессы, мероприятия, события, проверяются связи отдельных лиц и организаций с разного рода информационными ресурсами, определяется степень участия тех или иных лиц в действиях, которые могут нанести вред организации.

Проводимые исследования рассматриваются как с точки зрения тактических, так и с точки зрения стратегических целей.

Современный подход к анализу деятельности предприятия базируется не только на получении непосредственной информации о его работе, о протекающих на предприятии процессах, на изучении документации, наблюдении за работой предприятия, но и на серьезном исследовании открытых источников информации [7; 8]. С каж-

[1] Кандидат технических наук, доцент, доцент НОУ ВПО «Российский новый университет».
[2] Кандидат филологических наук, ИПИ РАН.
[3] Кандидат технических наук, ИПИ РАН.

дым годом всё большее влияние на этот процесс оказывает Интернет. Здесь большую трудность составляет фильтрация информации, ее аналитический разбор, формирование заключений по результатам анализа вследствие громадных объемов информационных ресурсов, представленных в открытых источниках. Особое значение приобретает информация, выложенная в Интернете, потому что в данном случае возможна автоматическая обработка информации, что было недоступно, пока информационные ресурсы не были переведены в электронный формат.

Для сокращения пространства поиска релевантной информации формируются ассоциативные портреты предметной области (АППО), предметные словари и тезаурусы [1; 2], образующие базу знаний. Это позволяет в значительной степени идентифицировать определенные объекты и процессы предметной области, существенно повысить полезность извлекаемой из Интернета информации. Для этого извлекаемые из Интернета тексты обрабатываются лингвистическим процессором.

В результате обработки больших массивов данных ассоциативные портреты предметной области, тезаурусы и предметные словари постоянно пополняются новыми данными, что позволяет улучшить качество фильтрации текстов из Интернета, повысить достоверность результатов аналитической обработки [3].

Для проведения аналитических исследований на предприятии может быть создана группа аналитиков, которая анализирует результаты автоматической обработки информации из следующих источников:

– сведения из Интернета;
– материалы из различных баз данных;
– материалы аналитических центров;
– электронные документы предприятия.

Существуют значительные проблемы при автоматической обработке данных из открытых источников:

– громадные объемы информации, спама в том числе;
– сложность идентификации конкретных объектов, процессов, ситуаций на основе анализа и сопоставления информации из различных источников;
– недостоверность данных;
– явная и неявная дезинформация;
– неполнота данных;
– постоянная изменчивость среды, ситуаций;
– историческая изменчивость, устаревание данных и т.д.

В результате постоянного совершенствования механизма распознавания и идентификации объектов и процессов, а также в силу накопления знаний, которые могут выверяться экспертами, эффективность процесса определения утечек информации и выявления действий отдельных субъектов, направленных против организации, может быть существенно повышена [5].

Одной из задач любого предприятия является задача формирования положительного имиджа предприятия. Для этого могут быть использованы результаты анализа внешнего и внутреннего окружения предприятия с целью формирования целенаправленных воздействий для коррекции имиджа.

Репутация компании представляет собой некую оценку группы индивидуумов о деятельности компании, группы людей или отдельного человека на основе определенных критериев. Тем более, что прослеживается четкая зависимость капитала компании от ее репутации. Понятия Goodwill – деловая репутация или Reputational Capital – репутационный капитал имеют прямое отношение к финансовому состоянию компании. В рамках обнаружения негативного контента о компании в Интернете и сведения к минимуму его влияния на компанию существуют фирмы, предлагающие услуги по управлению репутацией организации в Интернете (например, Online Reputation Management – ORM, Search Engine Reputation Management – SERM).

Обзор средств лингвистического, семантического и статистического анализа текстов

В настоящее время на рынке представлено достаточно большое количество различных средств для анализа текстов. Системы для исследования Интернета называют текстово-аналитическими, или процессорами сбора данных. Существует достаточно большое количество фирм, которые предлагают средства анализа текстов в Интернете, направленные на выявление утечек информации, анализ репутации фирмы и т.п. Ниже представлено описание некоторых из этих средств.

InfoTracer

Это программный пакет, представляющий собой инструментальное средство для осуществления поиска информации в сети Интернет. Эта фирма является членом союза Private Investigator Union. В рамках своего продукта предлагаются несколько видов поиска, таких, как Comprehensive Background Report, Criminal Records Search, People Search, Property Search, Email Search, Company Background Search, и других. Наиболее интересной является компонента системы по поиску преступников, определению их криминального окружения, адресов пребывания (текущих и прошлых), их возможных родственников, свя-

зей с криминальным миром. Кроме этого возможен поиск информации о наличии объектов собственности у преступников, информации в социальных сетях, возможно определение регистрации в различных учреждениях, поиск дат рождения, смерти, информации об оформлении супружества и расторжении брака, текущем статусе, о гражданстве и т.д.

Taiga (Noemic)

Taiga – это французская разработка для извлечения информации из различных баз данных, доступных в Интернете, из различных новостных сообщений, из публикаций трудов научных конференций и т.д. Эта система позволяет собирать информацию о наиболее обещающих научных разработках и областях, в которых конкуренция еще не достигла широких масштабов. Система анализирует не только текст, но также и рисунки, диаграммы, графики. Она собирает любую требуемую информацию, как только она появляется в Интернете. Информация о связях между двумя странами может быть обнаружена всего за несколько часов.

Acetic-company

Предлагается программное обеспечение Tropes для высокоскоростного анализа текстов. Его разработка основывается на семантической классификации с использованием ключевых слов, которые, в свою очередь, могут автоматически извлекаться из текстов естественного языка. В системе проводится качественный лингвистический анализ текстов.

Данная система может определять контекст, темы, главных действующих лиц при анализе ситуации. Она также позволяет извлекать объекты и субъекты, время и место действия, окружение и возможные цели при развитии ситуации. В системе реализован хронологический анализ текстов.

Менеджер онтологии естественного языка системы построен на основе семантических сетей, используются технологии анализа текстов естественного языка с использованием готовых и расширяемых классификаторов.

DuckDuckGo

Это машина поиска, которая составляет заключения в результате анализа текстов, выделяет категории, двусмысленности, определят оригинальные сайты как источники информации. Она может быть использована для определения людей, мест, вещей, слов и концепций. В системе возможно установление прямых связей с различными сервисами, реализован поиск релевантной информации, связанной с конкретными ресурсами в Интернете.

Amtera-company

Деятельность этой компании направлена на построение интеллектуальных систем, основанных на семантических представлениях и так называемых технологиях Big Data (аналитическая обработка больших объемов информации). Особенности системы: автоматический семантический анализ продуктов различных компаний, анализ тенденций развития ситуации на рынке (типа “Trend”-анализ), анализ потребления, широкие возможности настройки и поддержки продукта, семантический поиск (поиск по смыслу, не по ключевым словам).

Актуальность задачи аналитической разведки не вызывает сомнений. Большое количество коллективов занимается подобными вопросами. Использование методов семантического анализа практикуется во всем мире, что позволяет в существенной степени обезопасить предприятия от внешних и внутренних угроз, управлять репутацией организации в средствах массовой информации.

Семантическая обработка текстовой информации. Извлечение объектов, процессов, выделение отношений между объектами

В настоящее время для формализации знаний о предметной области и построении структуры бизнес-процессов активно используются различные методы представления знаний [4]. В данной статье рассматривается подход формирования знаний о предметной области в виде семантической сети, которая представляется в виде вершин и отношений между ними [9]. Обычно в виде вершин представляются некие объекты (субъекты) предметной области, а в виде отношений – действия, в которые данные объекты вовлечены, или связи между этими объектами. Семантическая сеть описывается в виде фрагментов [10–14]. Ниже представлены фрагменты семантической сети, описывающие объекты и связи между ними:

$$R1(A1,A2/N1)\ R2(A3,A4/N2)\ R3(A5,A6/N3). \quad (1)$$

Здесь: R1, R2, R3 – имена отношений [1];

N1, N2, N3 – имена фрагментов (N1 – имя всего фрагмента R1(A1, A2); при этом оно может опускаться, если этого не требуется для обработки специально); A1 – A6 – имена объектов.

Например, в результате анализа информации в Интернете был извлечен фрагмент текста:

«Бондарев принял участие в проекте по закупке оборудования для предприятия Nogos». Для удобства опущены имя и отчество. Этой фразе будет соответствовать фрагмент семантической сети:

Закупка оборудования (Nogos, Бондарев) (2)

При этом известно, что Бондарев не работает на предприятии Nogos и официально не участвует в проекте по закупке оборудования. Но на предприятии Nogos работает Кравцова, которая приходится сестрой Бондареву. В данном случае в систему должна быть помещена следующая информация в виде семантической сети (или эта сеть должна быть построена на основе анализа текстовой информации предприятия):

Сестра (Кравцова, Бондарев) (3)
Работник (Nogos, Кравцова). (4)

Данному множеству предикатов соответствует следующий фрагмент семантической сети (схема 1):

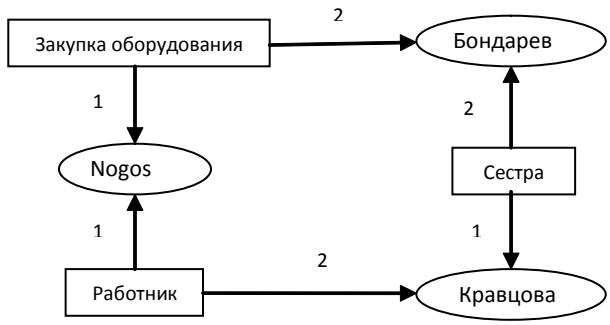

Схема 1

В результате анализа данной ситуации может быть сделан вывод, что работник предприятия Nogos Кравцова использовала свои родственные связи либо для продвижения продукта, либо для отмывания денег (возврата денег в виде «отката» за предоставление исполнителю по проекту заказа на закупку оборудования).

В данном случае можно говорить об использовании окрестностного подхода при анализе знаний. Рассматриваются окрестности объектов (субъектов) или вершин в рамках фрагмента семантической сети. Окрестность определяется наличием «близких» семантических связей. Различают окрестности первого, второго и т.д. порядка. Например, для объекта «Кравцова» господин Бондарев (ее брат) попадает в окрестность первого порядка, т.е. они связаны непосредственно через один из элементов семантической сети. В данном случае посредством фрагмента «Сестра» (Кравцова, Бондарев). Одновременно эти данные соотносятся с построенной в системе онтологией, которая содержит информацию о родственных связях (сестра, брат), классах объектов (класс людей). Кроме этого, если в системе уже присутствует описанная выше информация (о Бондареве, Кравцовой, их родственных отношениях), то происходит идентификация найденных, к примеру, в Интернете объектов с уже включенными в онтологию, что в существенной степени упрощает распознавание ситуации.

Описанный случай можно считать идеальным, иначе говоря, в данной ситуации не возникает трудностей с выводами. Но возможна более сложная ситуация, когда анализ семантической близости делается не только по родственным связям, а по опосредованным связям [6]. Рассмотрим следующий пример, внеся изменения в ситуацию:

«Петров принял участие в проекте по закупке оборудования для предприятия Nogos». При этом у нас нет прямой информации о связях господина Петрова. В случае если существует информация в Интернете, в социальных сетях о том, что Петров является другом Бондарева, тогда будут построены следующие фрагменты семантической сети:

Сестра (Кравцова, Бондарев) (3)
Работник (Nogos, Кравцова) (4)
Закупка оборудования (Nogos, Петров) (5)
Друг (Петров, Бондарев). (6)

В результате мы имеем следующее. Произошла утечка информации с предприятия; возможно, в этом случае предприятие понесло дополнительные убытки в виде тех же «откатов» или завышенной цены на оборудование. Ситуация описывается уже окрестностью второго порядка, когда отсутствуют прямые связи между объектами. В качестве фрагмента-посредника выступает предикат «Друг».

Данному множеству предикатов соответствует следующий фрагмент семантической сети (схема 2):

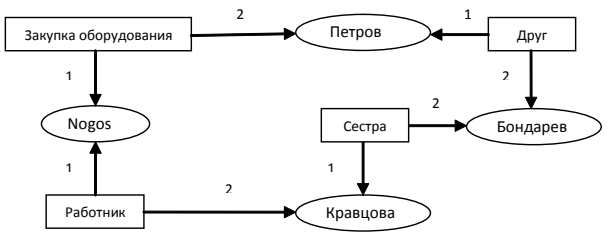

Схема 2

Окрестности вершин могут строиться по различным основаниям, когда не прослеживается непосредственная связь между объектами. Мож-

но определить наличие или отсутствие связи между объектами по ряду признаков, например:

– нахождение объектов в одном и том же месте (несколько раз);

– нахождение объектов в одном и том же месте в одно и то же время (несколько раз);

– обучение в одном и том же заведении;

– вхождение в одну и ту же организацию (клуб, автомобиль);

– похожесть интересов;

– локализация по месту жительства или работы (в широком смысле).

Данная ситуация может быть усилена дополнительной информацией, найденной в Интернете, если два объекта находятся в одном и том же месте в одно и то же время, и может быть описана фрагментом типа:

Место (Место, Объект, Месяц, Год). (7)

Например, если будет найден текст типа «Петров провел отпуск в июне 2015 в г. Сочи» и одновременно будет найден следующий текст «Бондарев обычно отдыхает в Сочи в августе», то к уже построенным фрагментам (исключая фрагмент Друг (Петров, Бондарев)) будут добавлены следующие фрагменты:

Место (Сочи, Петров, июнь, _) (8)

Место(Сочи, Бондарев, июнь, 2015). (9)

Это даст возможность судить о том, что эти два индивидуума могут быть связаны, в случае если Петров принимает участие в проекте закупки оборудования (см. выше). В результате будет построена следующая семантическая сеть (в данном фрагменте представлена обобщенная информация из онтологии данной предметной области – схема 3).

Во фрагменте семантической сети представлена информация из онтологии, а именно: «месяц», «год», «город», которая не была описана выше в вербальном виде. Знак «∈» определяет принадлежность, например объект «Сочи» принадлежит классу «Город», объект «июнь» принадлежит классу «Месяц», а объект «2015» принадлежит классу «Год».

Заключение

Рассматриваемый в данной статье подход позволяет не только сделать работу организации более безопасной, но и определить наиболее актуальные и перспективные направления ее развития.

Интерес представляют не только люди, которые могут нанести ущерб бизнесу, но и возможности, открывающиеся в результате изменения окружающей обстановки. Выявляются интересные новые направления развития, выделяются фирмы, субъекты, заинтересованные в конкретных сферах деятельности, анализируются связи между фирмами и направлениями развития, вычисляются наиболее приоритетные виды деятельности.

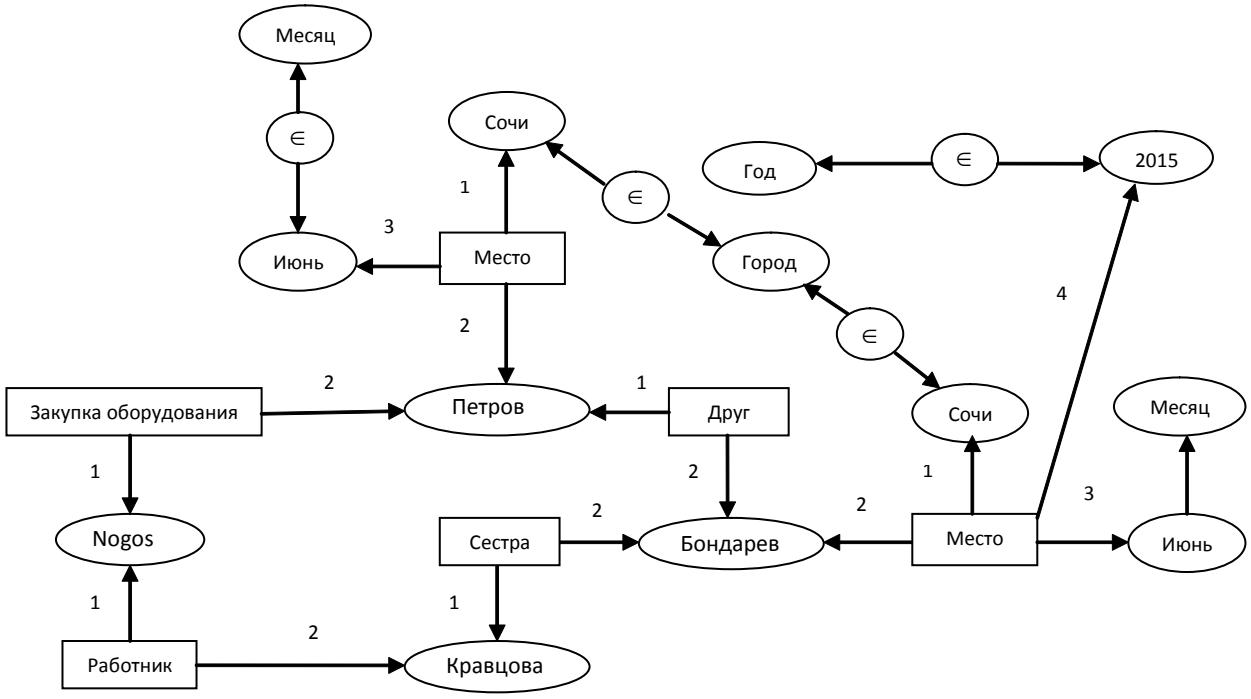

Схема 3

В результате анализа выявляется растущий к деятельности организаций интерес в прессе, выделяются стратегические цели развития предприятия.

Постоянный мониторинг открытых источников позволяет получить преимущества в следующих аспектах:

– анализ возможных рисков и возможностей развития;

– разработка плана упреждающих действий для борьбы с конкурентами;

– выявление новых направлений развития, появление новых конкурентов;

– анализ реноме компании в сообществе;

– определение каналов утечки информации;

– формирование положительного имиджа компании;

– выявление союзников.

Методика интегрального семантического анализа текстового пространства и извлечения значимых словосочетаний и опорных языковых структур для автоматического построения концептуальных моделей фокусных предметных областей является новой и позволяет эффективно строить и развивать семантико-ориентированные системы аналитической обработки знаний на основе интегрального подхода, включающего экспертные лингвистические эвристики в сочетании с методами машинного обучения, корпусной статистики, что имеет практическое значение для реализации эффективных систем поддержки принятия аналитических решений в сфере бизнеса и других предметных областей.

Литература

1. Zolotarev, O., Charnine, M., Matskevich, A. Conceptual Business Process Structuring by Extracting Knowledge from Natural Language Texts // Proceedings of the 2014 International Conference on Artificial Intelligence (ICAI 2014). – Vol. I. – WORLDCOMP'14, July 21–24, 2014. – Las Vegas Nevada, USA: CSREA Press. – Pp. 82–87.

2. Charnine, M., Somin, N., Nikolaev V. Conceptual Text Generation Based on Key Phrases // Proceedings of the 2014 International Conference on Artificial Intelligence (ICAI 2014). – Vol. I. – WORLDCOMP'14, July 21–24, 2014. – Las Vegas Nevada, USA: CSREA Press. – Pp. 639–643.

3. Михеев М.Ю., Сомин Н.В., Галина И.В., Золотарев О.В., Козеренко Е.Б., Морозова Ю.И., Шарнин М.М. Фальштексты: классификация и методы опознания текстовых имитаций и документов с подменой авторства // Информатика и ее применения. – 2014. – Том 8. – Выпуск 4. – М. : РАН.

4. Золотарев О.В., Шарнин М.М., Козеренко Е.Б. Построение моделей бизнес-процессов на основе выделения процессов, объектов и их связей из текстов естественного языка // Физико-техническая информатика (СРТ2014) : труды Международной конференции (Ларнака, Кипр, 11–18 мая 2014). – Протвино : ИФТИ, 2014.

5. Золотарев О.В., Козеренко Е.Б., Шарнин М.М. Принципы построения моделей бизнес-процессов предметной области на основе обработки текстов естественного языка // Вестник РосНОУ. – 2014. – № 4. – С. 82–88.

6. Золотарев О.В. Процессный подход к управлению в проектах внедрения корпоративных информационных систем // Вестник РосНОУ. – 2014. – № 4. – С. 89–92.

7. Золотарев О.В. Методы выделения процессов, объектов, отношений из текстов естественного языка // Проблемы безопасности российского общества. – Смоленск : Свиток, 2014.

8. Золотарев О.В. Инновационные решения в формировании функциональной структуры предметной области // Вестник РосНОУ. – 2013. – № 4. – С. 82–84.

9. Золотарев О.В. Управление в проектах внедрения распределенных корпоративных информационных систем // Вестник РосНОУ. – 2012. – № 4. – С. 78–80.

10. Морозова Ю.И., Козеренко Е.Б., Шарнин М.М. Методика извлечения пословных переводных соответствий из параллельных текстов с применением моделей дистрибутивной семантики // Системы и средства информатики. – 2014. – Т. 24. – Вып. 2. – С. 131–142.

11. Сомин Н.В., Шарнин М.М. Использование хеш-функций для повышения скорости морфологического анализа русских текстов // Системы и средства информатики. – 2014. – Т. 24. – Вып. 3. – С. 204–217.

12. Козеренко Е.Б. Интегральное моделирование языковых структур в лингвистических процессорах систем обработки знаний и машинного перевода // Информатика и ее применения. – 2014. – Т. 8. – Вып. 1. – С. 89–98.

13. Мацкевич А.Г. Декларативные структуры знаний в проблемно-ориентированных системах искусственного интеллекта // Информатика и ее применения. – 2014. – Т. 8. – Вып. 2. – С. 122–129.

14. Морозова Ю.И., Козеренко Е.Б., Будзко В.И., Кузнецов К.И., Шарнин М.М. Семантическая структуризация текстовых знаний для систем аналитического мониторинга больших объемов информации в социальной сфере // Системы высокой доступности. – 2014. – Т. 10. – № 3. – С. 21–35.