# ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ ПРИМЕНЕНИЯ СЕМАНТИЧЕСКИ ОРИЕНТИРОВАННЫХ МЕТОДОВ КЛАССИФИКАЦИИ ЗАДАЧИ OPINION MINING

**Нина Хайрова, Наталья Шаронова**

***Аннотация**: Как правило, для того чтобы принять верное бизнес решение, необходимо прогнозировать изменения на рынке и в обществе в целом. Интеллектуальный анализ знаний, содержащихся в текстах Web-ресурсов пользовательского контента, сегодня в основном, осуществляется с помощью «ручной» работы аналитиков, скорость работы которых не соответствует темпам социально-экономических процессов общества. Хорошим дополнительным средством решения проблемы интеллектуального анализа слабоструктурированных текстов Web-ресурсов может стать отслеживание общественного мнения по отношению к тому или иному объекту или субъекту анализа, с использованием методов sentiment классификации. В работе предлагается система, реализующая методы opinion mining, позволяющая извлекать полезные знания из пользовательского Web-контента, использующая информационные технологии семантической классификации и идентификации многоязыковых семантических эквивалентов, и базирующаяся на специально-структурированном тезаурусе. Система использует механизм классификации мнений, позволяющий добавлять оценочные правила, представляющие позитивные, негативные и нейтральные оценки характеристик слов. В качестве основания классификации слабоструктурированной информации используются отношения содержащихся в ней аргументов к исследуемому объекту. Соответствие полного текста классу позитивного или негативного мнения определяется статистико-семантическими методами, основанными на атрибутивной положительной или отрицательной оценке объектов (субъектов) предложений. Кроме того, в работе показана специальная структура тезауруса, с явно заданными отношениями между концептами, используемыми для решения задач Opinion Mining, и идентификации семантических эквивалентов. Экспериментальные результаты на полнотекстовой базе пользовательского Web-контента текстов на английском и немецком языках показали достаточно высокую эффективность использования системы.*

***Ключевые слова**: интеллектуальный анализ знаний, opinion mining, семантическая классификация, онтологии, тезаурусы.*

***ACM Classification Keywords** : H.3.3 .Information Search and Retrieval*

## Введение

Для того чтобы принять верное бизнес решение, необходимо прогнозировать изменения на рынке и в обществе в целом. Правильный прогноз во многом базируется на отслеживании общественного мнения по отношению к тому или иному объекту инвестирования. Сегодня для правильного принятия финансово значимых экономических решений необходимо анализировать большое количество информации,

находящейся как в специализированных экономических источниках (например, информационные источники рейтинговых агентств Fitch, Moody's, Standard@Poor's, A.M.Best), так и в и источниках общего характера.

Традиционно инвесторы, принимая решения на фондовом рынке о входе на рынок (покупка акций) или выходе из него (продажа акций), обычно используют технический анализ, основанный на анализе временных рядов цен акций, с привлечением различных индикаторов, а также методы фундаментального анализа, базирующиеся на отчетной информации о текущем финансовом состоянии компаний эмитентов акций.

Однако, в связи с тем, что объемы неструктурированной информации увеличиваются с каждым годом, и сегодня, по данным агентства Gartner, неструктурированные документы составляют более 80% корпоративных данных, а количество внешних источников (интернет-ресурсов, блогов, форумов, СМИ) исчисляется миллионами, мощным дополнительным средством прогнозирования ожидаемых изменений на фондовых рынках может стать информация, извлекаемая из текстов соответствующей проблематики. Интеллектуальный анализ знаний, содержащихся в текстах Web-ресурсов, электронных журналах, электронной корреспонденции, социальных сетях и другом пользовательском контенте (user generated content), на сегодня, по большей части, осуществляется с помощью «ручной» работы аналитиков, скорость работы которых не соответствует интенсивности работы фондового рынка. Поэтому все чаще основным вопросом эффективности принятия бизнес решений становится скорость извлечения информации и постоянный ее мониторинг.

## Актуальность исследования

Различные системы, использующие Text Mining, могут предоставить бизнес аналитику возможность работать с большими объемами исходных данных за счет автоматизации процесса извлечения нужной информации [Gupta, 2009]. Но, несмотря на наличие большого количества подобных систем, к настоящему моменту нет систем, осуществляющих полную автоматическую обработку слабоструктурированных текстов, представленных в корпоративных и глобальных информационных системах. Это связано, прежде всего, с тем, что для принятия бизнес решения необходимо быстрое получение полной и релевантной информации, а коэффициенты полноты и точности существующих гипертекстовых систем составляют в среднем соответственно 0,9, и 0,8, а документальных систем - даже 0,7 [Gupta, 2009], что представляется довольно низким для принятия правильных бизнес решений.

Используемые технологии поиска и классификации слабоструктурированной информации базируются, в своем большинстве, на традиционных статистико-вероятностных подходах, не используя или слабо используя семантический, т.е. смысловой анализ [Нгок, 2012]. Это поиск по ключевым словам, основанный на техниках двоичного поиска, ранжированного поиска, вероятностных моделях поиска и т. д. Такой поиск не учитывает ни семантические связи между понятиями, ни наличие семантических эквивалентов слов и словосочетаний [Hany, 2011].

Информационное обеспечение процессов принятия бизнес решений может стать более качественным благодаря использованию, при решении общей задачи получения и обработки информации, моделей и методов Opinion Mining [Ding, 2008], в которых, в качестве основания классификации слабоструктурированной информации пользовательского контента, используется отношение (например,

положительное или отрицательное, поддерживающее или отклоняющее) содержащихся в них аргументов к исследуемому объекту [Esuli, 2006].

Дальнейшее развитие методов Text Mining в направлении информационного обеспечения бизнес решений сдерживается не только отсутствием адекватных моделей и методов семантического анализа, семантического поиска и смысловой классификации, но и отсутствием стройной системы лингвистического обеспечения [Оробинская, 2010]. Недостаточность онтологий и тезаурусов широких предметных областей приводит к невозможности учитывать особенности смысловой организации текстов по конкретной тематике, в частности, особенности смысловой организации экономических текстов.

## Общая постановка задачи

Для получения полной информационной картины по определенному объекту (например, акции) или субъекту (например, банк) необходимо использовать методы Opinion Mining, представляющие инструменты для автоматического извлечения из текстов «субъективной» информации, используемой для автоматической оценки (позитивной, негативной, нейтральной) новостных событий, продуктов, персоналий, организаций, стран мира и т.д., поступающих в режиме реального времени из сообщений электронных средств массовой информации, сообщений блоггеров, дискуссионных форумов и т.д.

Инструменты Opinion mining включены во многие системы, например, Salience Engine (Lexalytics, Inc., Boston, USA); SentiMetrix (University of Maryland, Institute for Advanced Computer Sciences), Twitter Sentiment (Stanford University), J.D.Power Text Analyst (J.D. Power and Associates, USA), RavenPack News Analyst (RavenPack International S.L.) и др. Для решения данной задачи используются алгоритмы опорных векторов и нейронных сетей, семантические сети, ассоциативные и оценочные правила, наивный байесовский метод, CRF модели и др. [Kobayashi, 2007]. Однако реального практически используемого результата, позволяющего включить данную технологию в систему информационного обеспечения процессов бизнес решений, на сегодняшний день не существует.

Задача осложняется наличием такой неотъемлемой составляющей естественных языков, как синонимия. И хотя сегодня существуют методы автоматического выделения смысловых эквивалентов из текста, а также синонимические связи учитываются онтологиями и тезаурусами, в практической работе для учета семантической эквивалентности пользователю часто приходится повторять запросы с каждым из «вручную» определенных смысловых эквивалентов термина.

Таким образом, на основе проведенного анализа существующих моделей и методов решения задачи Opinion mining в работе предлагается рассмотреть задачу Opinion Mining и как часть общей задачи получения и обработки информации, и как задачу автоматизированного построения классификатора слабоструктурированных текстов, представляющих так называемый «user generated content». Кроме того, для получения релевантного результата при осуществлении Social Media Monitoring необходимо учесть поступление информации на различных языках и использование в текстовых сообщениях семантических (смысловых) эквивалентов.

В настоящей работе категория мнения рассматривается как субъективные заявления людей, отражающих настроения или представления о сущностях или событиях, а информационная технология решения задачи Opinion Mining рассматривается, как поэтапное решение следующих задач:

- идентификации объекта или субъекта мнения;
- идентификации сущностей, выражающих мнения;
- идентификации полярности мнения;
- выявления семантических эквивалентов.

## Структурное описание модели

Проведенный обзор показывает, что 28% информации, значимой для принятия бизнес решений, находится в слабоструктурированных текстах. Информационное обеспечение, необходимое для обработки данной информации, должно включать в себя как информационный поиск, содержащий мультилингвистическую семантическую задачу, напрямую связанную с релевантностью результатов, — выявления семантических эквивалентов, так и методы Text Mining. Метод Opinion Mining, позволяющий осуществлять отнесение текстов, по мнению авторов, к определенной проблеме, появился сравнительно недавно, и во многом использует подходы Text Mining и Information Retrieval (рис. 1).

Тексты экономически значимых документов, в подавляющем большинстве, оформлены в официально-деловом стиле, лексика которого характерна логической связностью и насыщенностью экономическими и специальными общенаучными терминами. Такого рода термины представляют собой эмоционально-нейтральные слова (словосочетания), передающие название точно определяемых понятий, относящихся к определенной области науки. Использование в текстах такой лексики дает возможность наиболее четко, точно и экономно излагать содержание данного предмета и обеспечивает правильное понимание существа трактуемого вопроса. В экономически значимых текстах, которые и являются информационной базой для принятия серьезных бизнес решений, термины несут основную семантическую нагрузку, занимая главное место среди прочих общелитературных и служебных слов.

Такое использование экономически значимой научной и «околонаучной» терминологии позволяет довольно эффективно использовать для анализа подобных текстов электронные толковые словари, тезаурусы, онтологии соответствующих предметных областей, на которых обоснованно должны базироваться задачи Text Mining, Opinion Mining и Information Retrieval (см. рис. 1)

Сегодня задачи выявления семантических эквивалентов и Opinion mining решаются изолированно, с использованием, в подавляющем большинстве, статистических и, в редких случаях, комбинированных методов. Использование семантических методов хотя и позволяет получать более точные результаты, но развитие подобных моделей, с использованием нейронных или семантических сетей, представляют довольно сложный и трудоемкий процесс.

Предлагается для решения задачи Opinion mining использовать семантически ориентированные методы классификации (sentiment classification), с добавлением оценочных правил, представляющих позитивные, негативные и нейтральные оценки характеристик слов (рис. 2). Используется комбинированный семантико-статистический метод, семантическая составляющая в котором основывается на использовании положительной и отрицательной связи аргументов (слов и словосочетаний) с объектом анализа [Osgood, 1957].

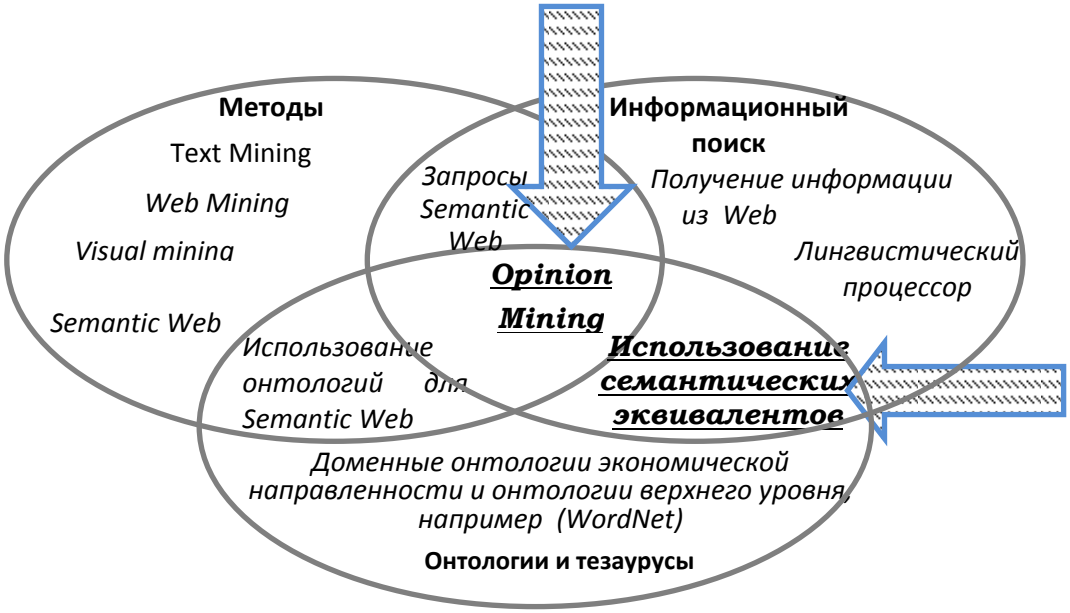


Рис. 1. Общая структура формирования информационного обеспечения задачи принятия бизнес решений при обработке слабоструктурированной экономически значимой информации.

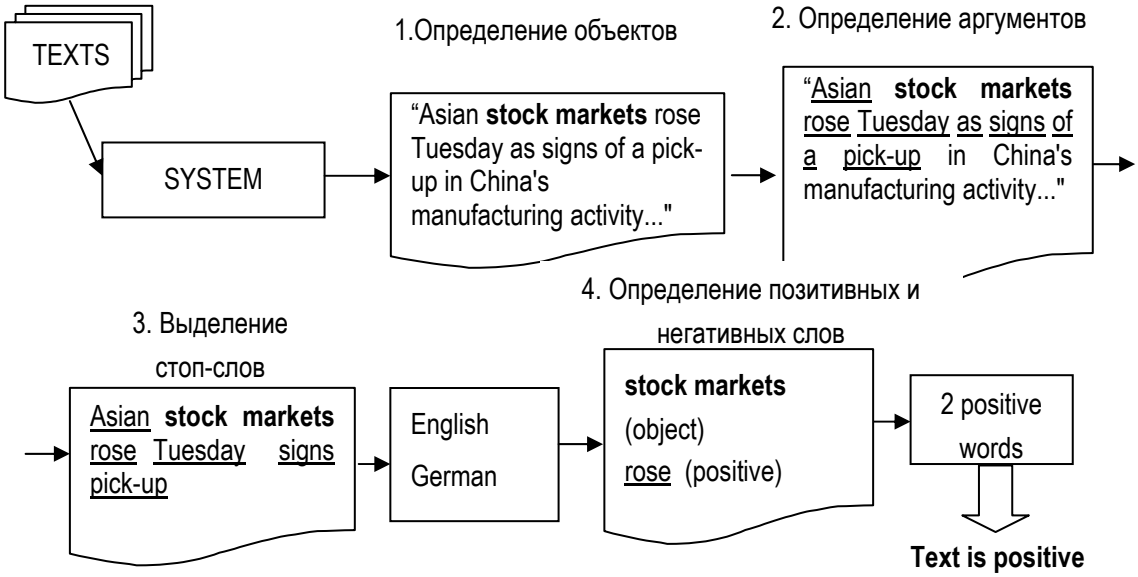


Рис. 2 Структурная схема семантически-ориентированной задачи классификации Opinion mining

Для выявления семантической силы положительного или отрицательного мнения по отношению к тексту определяется атрибутивная оценка объектов (субъектов) небольших фрагментов текста (фраз или предложений). Определение объектов мнений и аргументов, выражающих классификационные признаки мнений, а также их эквивалентов базируется на предварительном осуществлении предлингвистического анализа, стеминга и формального синтаксического анализа [Хайрова, 2010], в результате которых осуществляется нормализация лексем. Нормализация, или приведение лексем к канонической форме, осуществляется для устранения несущественных различий между последовательностями символов, в результате чего множество словоформ предложения преобразуется в список терминов.

На следующем этапе для повышения степени релевантности результата используемой технологии обрабатываются отрицательные слова и стоп слова языка. Отрицательные слова представляют собой четко очерченное множество шумовых слов, подобных словам 'I', 'me', itself', 'being', 'after', 'about', 'can' и т.п. английского языка или подобные слова любого естественного языка, которые исключаются из списка аргументов.

Обработка отрицаний предполагает использование шаблона регулярных выражений для таких форм как 'can't', 'mustn't' и др. и отдельную обработку слова 'not':

if ('not' before arg (pozitive)) $\Rightarrow$arg (negative) and

if ('not' before arg (negative)) $\Rightarrow$arg (pozitive)

Для определения позитивности (аналогично негативности) предложений текста, по отношению к объекту принимаемого бизнес решения, определяется принадлежность нормализованных лексем слов, находящихся на расстоянии *k* от объекта, аргументами или характеристиками данного слова. Основываясь на работе [Osgood, 1957], в качестве оптимального значения выбрано $k \leq 4$.

Разработанный тезаурус, базирующийся на свободно распространяемом словаре английских слов Minqing Hu and Bing Liu [Bing, 2005], включает 6800 позитивных и негативных английских существительных, глаголов (3 основных используемых формы: инфинитив, past и 3-я форма единственного числа), прилагательных и наречий.

Целостная оценка мнения текста по анализируемому объекту вычисляется статистически с помощью определения веса в тексте соответственно позитивных или негативных, по отношению к запрашиваемому объекту, предложений:

$$O_p = \frac{\sum_{i=1}^{k} S_i^p}{\sum_{j=1}^{n} S_j} \qquad (1)$$

где $O_p$ — вес позитивной оценки текста, $S_j$ — любые предложения текста, содержащие объект исследования, n — число фраз или предложений текста, содержащих объект исследования, $S_i$ — предложения текста, положительный аргумент в которых находится на расстоянии <= 4 от исследуемого объекта оценки, k — число таких фраз или предложений в анализируемом тексте. Аналогичным образом вычисляется вес негативной оценки текста.

## Структура используемого тезауруса

В наиболее полных открытых мультиязычных онтологиях типа STW Thesaurus for Economics (by Information Center for Economics), WordNet (by Princeton University), Finance ontology (by Eddy Vanderlinden), Ruthes (Russian Center for information researches), Gemet (by EEA and Eionet), SentiWordNet используются отношения, покрывающие довольно большой класс семантических задач: symmetrical, associativity, homonymy, synonymy, hyponymy, hyperonymy, measuring, symmetrical,association, equivalence, antonymy, troponymy, asymmetrical, value judgments [Соловьев, 2006].

Для решения задачи выявления характеристик текстов, передающих субъективное отношение говорящего (пишущего) к содержанию или адресату речи, с учетом возможности разноязычных высказываний с использованием смысловых эквивалентов, используется тезаурус, в котором явно определены следующие связи между концептами:

1. Синонимия (synonymy) — отношение, связывающее, термины, являющиеся синонимами
2. Эквиваленты (equivalence) — отношение, связывающее переводные эквиваленты одинаковых концептов, на разных языках.
3. Ценностное суждение (value judgments) — отношение между объектами и словами, которые описывают определенный эмоциональный оттенок объекта.
4. Антоним (antonymy) — отношение, связывающее слова и словосочетания с противоположным значением.

Под концептом понимается абстрактная группа, коллекция или выборка сущностей, которая может включать отдельных представителей, другие классы или и то, и другое. На рисунке 2 приведен структурный фрагмент тезауруса, для концепта «economic area» и возможных позитивных, негативных и нейтральных суждений для английских и немецких семантических эквивалентов. На рисунке представлены следующие отношения:

1. Synonymy (economic area⇒ economic geography, economic area⇒ industrial area);
2. Equivalence (economic area⇒ Wirtschaftsgebiet (de), economic area ⇒ Wirtschaftsraum (de), economic area ⇒Wirtschaftsregion (de));
3. Antonym (small ⇒ big, expensive⇒ cheap,…);
4. Value judgment: positive (economic area⇒ big); negative (economic area⇒ unproductive); neutral (economic area⇒interesting).

Онтология представляет собой формальное описание знаний в виде множества концептов в предметной области и связей между этими концептами, используемое для решения большого класса задач, в том числе онтология может быть использована для описания и вывода сущностей предметной области. Среди множества языков, используемых в настоящее время для описания онтологий (CycL, F-Logic, KL-ONE, LOOM, OCML, DAML-OIL, Web Ontology Language (OWL), RDF (Schema), SQL) наиболее популярными по-прежнему остаются OWL, RDF, SQL.

Информационное обеспечение задачи Opinion Mining, использующее мультиязыковые семантические эквиваленты, представлено тезаурусом, реализованным в виде реляционной базы данных. Выбор языка SQL обоснован возможностью описания определенных выше отношений между концептами на разных языках, простотой использования и наличием инструментов перехода от языка SQL к языку RDF и наоборот.

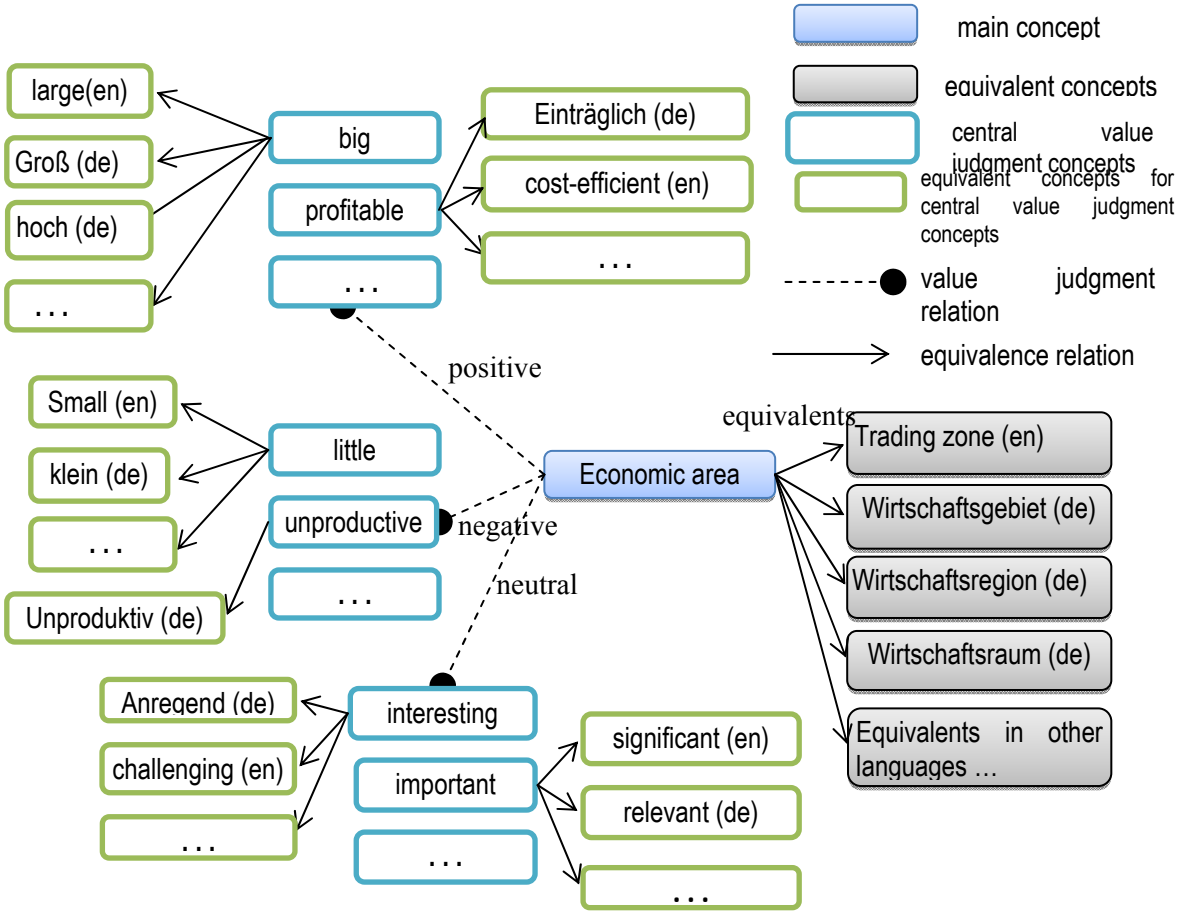


Рис. 2. Структурный фрагмент тезауруса, используемого для решения задачи Opinion Mining, с выделением семантических эквивалентов

## Программная реализация модели

Для тестирования алгоритма будет использован инструмент Natural Language Toolkit, обладающий преимуществом свободного распространения и включением большого количества встроенных функций обработки языков.

Для проверки работы алгоритма были протестированы несколько десятков текстов различной экономической направленности – RSS - рассылок на английском и немецком языках. В подавляющем большинстве случаев (87%) классификация текста рассмотренной системой, использующей средства opinion mining, совпадала с мнением эксперта о положительном (отрицательном) мнении автора текста об анализируемом субъекте (объекте).

## Выводы

Результатом данного исследования является разработка информационной технологии, позволяющей осуществлять классификацию текстов Web-ресурсов пользовательского контента с применением в качестве основания классификации извлеченных знаний о семантически ориентированном мнении, содержащемся в тексте по отношению к тому или иному объекту или субъекту анализа. Предложенная

технология включает в себя методы Opinion Mining, позволяющие автоматизировать извлечение из текста знаний об отношении автора (позитивном, негативном, нейтральном) к рассматриваемому предмету анализа и методы идентификации многоязычных семантических эквивалентов. Применение данных методов базируется на специально разработанном тезаурусе, структура которого в явном виде реализует отношения синонимии, эквивалентности, антонимии и ценностных суждений между концептами. Тезаурус, представляющий информационное обеспечение задачи Opinion Mining, использующей мультиязычные семантические эквиваленты, реализован в виде реляционной базы данных.

Программная реализация модели показала возможность ее использования при классификации английских и немецких текстов экономической направленности по отношению к определенному объекту или субъекту анализа.

**Литература:**

[Bing, 2005] Liu, Minqing Hu and Junsheng Cheng. "Opinion Observer: Analyzing and Comparing Opinions on the Web" Proceedings of the 14th international World Wide Web conference (WWW-2005), May 10-14, 2005, in Chiba, Japan.

[Ding, 2008] Ding Xiaowen, Liu Bing and Philip S Yu. . A Holistic Lexicon-Based Approach to Opinion Mining. Proceedings of the first ACM International Conference on Web search and Data Mining (WSDM'08), 2008.

[Esuli, 2006] Andrea Esuli, Fabrizio Sebastiani. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006), Genova, IT, 2006, pp. 417-422.

[Gupta, 2009] Vishal Gupta, Gurpreet S. Lehal A Survey of Text Mining Techniques and Applications / Journal of Emerging Technologies in Web Intelligence, Vol 1, No 1 (2009), 60-76, Aug 2009.

[Hany, 2011] Hany M. Harb, Khaled M. Fouad, Nagdy M. Nagdy Semantic Retrieval Approach for Web Documents / International Journal of Advanced Computer Science and Applications (JACSA), Vol. 2, No. 9, 2011 — pp.67-76.

[Kobayashi, 2007] Nozomi Kobayashi, Kentaro Inui and Matsumoto Yuji. Opinion Mining from Web Documents: Extraction and Structurization. Journal of Japanese society for artificial intelligence, Vol.22 No.2, special issue on data mining and statistical science, pages 227-238, 2007.

[Osgood, 1957] Osgood, Charles E., Suci, George J., and Tannenbaum, Percy H. 1957. The Measurement of Meaning. Urbana: university of Illinois Press.

[Нгок, 2012] Нгуен Ба Нгок, Тузовский А.Ф. Обзор подходов семантического поиска / Доклады ТУСУРа. Управление, вычислительная техника и информатика. № 2 (22), часть 2, 2010. — С. 234-237.

[Оробинская, 2010] Оробинская Е.А., Кочуева З.А. Технологии TEXT MINING: обзор методов и задач обработки смысловой информации // Информационные технологии / Вестник ХНТУ №2(38), 2010. с. 348-353.

[Соловьев, 2006] Соловьев В. Д., Добров Б.В., Иванов В.В., Лукашевич Н.В. Онтологии и тезаурусы. Учебное пособие. Казань, Москва. 2006. — 157с.

[Хайрова, 2010] Хайрова Н. Ф.. Тарловский В. А. Использование семантико-ориентированного лингвистического процессора для добывания новых знаний из потока документов корпоративной информационной системы / Вісник Національного технічного університету «ХПІ». Збірник наукових праць. Тематичний випуск «Системний аналіз, управління та інформаційні технології». — Х.: НТУ «ХПІ». — 2010. — № 67. — С. 132-138.

## Сведения об авторах

*Хайрова Нина – доцент кафедры интеллектуальных компьютерных систем Национального технического университета «Харьковский политехнический институт», ул. Фрунзе, 21, Харьков, 61002, Украина e-mail: nina_khajrova@yahoo.com*

*Научные интересы: искусственный интеллект, обработка знаний, автоматическая обработка текстов*

*Шаронова Наталья – профессор, заведующий кафедрой интеллектуальных компьютерных систем Национального технического университета «Харьковский политехнический институт», ул. Фрунзе, 21, Харьков, 61002, Украина e-mail:, nvsharonova@mail.ru*

*Научные интересы: искусственный интеллект, математическое моделирование, автоматизированные библиотечные системы*