УДК 519.52

# ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ТЕКСТОВ В СИСТЕМАХ МЕНЕДЖМЕНТА ЗНАНИЙ

**О. В. Шубкина**

ХНУРЭ, г. Харьков, Украина
olga.shubkina@gmail.com

Проведен анализ задач, стоящих перед разработчиками систем менеджмента знаний, а также анализ существующих методов интеллектуального анализа текстов в системах управления знаниями. Описан метод концептуального аннотирования для формирования семантических профилей документов, на основе которого разработана архитектура системы.

УПРАВЛЕНИЕ ЗНАНИЯМИ, ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ТЕКСТОВ, КОНЦЕПТУАЛЬНОЕ АННОТИРОВАНИЕ, СЕМАНТИЧЕСКИЙ ПРОФИЛЬ ДОКУМЕНТА

## Введение

На современном этапе социально-экономического развития человечества – этапе построения информационного общества – знание, его создание и управление им становится ключевым ресурсом развития мировой экономики. Деятельность как отдельных людей, так и организаций все в большей степени зависит от имеющихся у них знаний и способности эффективно их использовать.

Менеджмент или управление знаниями (Knowledge Management – KM) сегодня рассматривается как мощное конкурентное преимущество. Однако ни информационные технологии (ИТ), ни данные сами по себе не могут обеспечить конкурентного преимущества на долгосрочный период. Они могут быть достигнуты только «переводом» информации в ценные, смысловые руководства к действию. Таким образом, знание состоит в действии: в эффективном представлении данных и информационных ресурсов для принятия решений, а также в самом выполнении принятого решения [1].

Ресурсы знаний различаются в зависимости от отраслей индустрии и приложений, но, как правило, представлены в текстовом виде. Документы представляют собой ресурс знаний организации, в котором находится около 85% всей информации [2, 3]. Таким образом, возрастает роль извлечения знаний из различного вида текстовых источников, накопленных на разных этапах развития организации. Этот процесс является основной задачей Text Mining (TM) – интеллектуального анализа текстовой информации [4]. В управлении знаниями ТМ играет важную роль, потому что является механизмом выявления закономерностей, характерных элементов или свойств, которые могут использоваться в качестве метаданных документа, ключевых слов, аннотаций. Другая важная задача ТМ состоит в отнесении документа к некоторым категориям из заданной схемы их систематизации. Таким образом, с помощью ТМ можно обратить в свою пользу всю имеющуюся у организации информацию, опыт и квалификацию сотрудников с тем, чтобы повысить качество обслуживания клиентов и сократить время реакции на меняющиеся рыночные условия.

## 1. Постановка задач исследования

В рамках данной работы необходимо провести анализ задач, стоящих перед разработчиками систем КМ и возможности их решения с помощью интеллектуального анализа текстовой информации, предложить подход к алгоритмизации аннотирования текстовых документов на основе имеющейся концептуальной модели предметной области (ПрО). На основе анализа известных методов ТМ разработать метод концептуального аннотирования (КА) для создания профилей документов, а также сформировать обобщенную архитектуру системы КА на основе полученного алгоритма.

## 2. Основные задачи, решаемые с помощью систем КМ

Менеджмент знаний, несмотря на междисциплинарный характер данного направления исследований, в настоящее время наиболее активно развивается в искусственном интеллекте, предлагая прогрессивные решения, основанные на парадигме Semantic Web, по созданию систем корпоративной памяти предприятий, порталов знаний и другого [5]. К основным задачам КМ относят: достижение поставленных целей организации за счет роста интеллектуального капитала и эффективного его использования; повышение эффективности принимаемых решений; создание предпосылок для появления инноваций; повышение эффективности процессов проектирования. Основные задачи КМ решаются с помощью разработки корпоративных систем менеджмента или управления знаниями (СУЗ), реализуемых в настоящее время на основе парадигмы корпоративной памяти (КП). Основное предназначение КП состоит в накапливании, систематизации, управлении и совместном использовании профессиональными группами сотрудников всей необходимой и полезной информации в целях достижения конкурентного преимущества организации, эффективного и сво-

евременного решения текущих задач, исключении дублирования, противоречивости и потери знаний, накопленных в процессе жизнедеятельности организации. Таким образом, наиболее важной задачей, стоящей перед разработчиками СУЗ, является разработка подхода к проектированию КП, от которой зависит эффективность работы системы в целом.

Среди интеллектуальных информационных технологий, используемых для разработки СУЗ, значительную роль играет ТМ, целью которого является извлечение полезной информации из текстовых источников путем идентификации и исследования полезных интересных шаблонов. Поскольку большая часть информации как в корпоративных Intranet-сетях, так и в WEB-пространстве содержится в текстовом виде, технологии интеллектуальной обработки текстов помогают решать многие задачи КМ, и прежде всего, построение пространства знаний (по сути, КП) путем формирования адекватной онтологии ПрО на основе извлечения знаний из предметно-ориентированных текстовых коллекций, их структурирования и анализа. Несмотря на то, что существуют классические известные подходы, модели и языки описания знаний ПрО, в настоящее время de facto наиболее эффективным стал онтологический подход к представлению знаний, развиваемый в рамках нового направления искусственного интеллекта – онтологического инжиниринга (Ontology Engineering - OE). Именно онтология позволяет получить эффективное представление эксплицитной концептуальной модели для сложноструктурированных и слабоформализованных ПрО в гетерогенных распределенных пространствах, к которым относится и КП предприятия.

### 3. Современные средства ТМ

В настоящее время наиболее востребованы системы ТМ с максимально автоматизированными ETL-процессами структурирования контента (Extract, Transfer, Load — извлечение, преобразование, загрузка). Важной характеристикой такого рода систем является функция оперативного анализа информации, полученной по запросу для выбора дальнейшего направления исследования документов (автопилотирование направления исследования), выполняемая с помощью методов ТМ.

В таких системах используется двухфазная технология аналитической обработки. В первой фазе (ETL) производится автоматизированный анализ отдельных документов, структуризация их контента и формирование хранилищ исходной и аналитической информации. Во второй фазе (OLAP, Text Mining, Data Mining) — извлечение в оперативном режиме знаний из хранилища или из полученной по запросу подборки документов. К наиболее интересным системам аналитической обработки относятся ClearForest, Convera RetrievalWare, Hummingbird KM, IBM Text Miner, инструменты компании IQMen, Inxight Smart Discovery Extraction Server, Ontos Miner, Oracle Text, ODB-Text, TextAnalyst, инструменты компании Smartware, XANALYS Link Explorer, X-Files [6].

К наиболее актуальным направлениям извлечения знаний из текста на сегодняшний день относятся: аналитическая обработка фактов; ведение досье; извлечение и структурирование фактографической информации; поиск информации по запросам на естественном языке с использованием тезаурусов; направление поиска информации, объектов в хранилище документов, в подборке документов; аннотирование документов, построение дайджестов по объектам; проведение тематического анализа документов (кластеризация и рубрицирование); построение и динамический анализ семантической структуры текстов; выделение ключевых тем и информационных объектов; определение общей и объектной тональности сообщений; исследование частотных характеристик текстов.

### 4. Аннотации как способ извлечения знаний из текстов

Обработка текстов подразумевает несколько последовательных этапов: на первом происходит нормализация слов с учетом морфологии языка; на втором – семантический анализ текста, когда уточняется конкретный смысл слова в зависимости от контекста. Затем строится семантический образ исходного документа, на основе которого делаются интеллектуальные запросы.

Классический подход к аннотированию документов предусматривает определение в тексте основных тем, позволяет выделить предложения, в которых тематика документа представлена наиболее ярко (представлены все основные тематические узлы). Данные предложения и образуют аннотацию текста.

Потребность в автоматическом составлении аннотаций для поисковых машин и каталогов довольно высока и со временем увеличивается. Держатели каналов вещания, списков рассылки новостей и корпоративных систем документооборота также начинают испытывать интерес к автоматизации этого процесса. Зачастую не только требование дать возможность быстрого просмотра содержания, но и желание не нарушать авторские права заставляет заменять полные тексты документов аннотациями. Пользователю также может пригодиться возможность быстро просматривать большие объемы документов, используя их краткие аннотации.

Следует отметить, что фактически во всех известных системах машинное аннотирование является экстрагированием – программа не «пересказывает» смысл текста, а просто извлекает из него те фрагменты, которые считает важными, и объединяет их в аннотацию. Важность конкретного предложения определяется по различным параметрам, в частности по так называемым маркерам важности (например, «в заклю-

чение нужно сказать, что...»), количеству содержательных слов в нем и так далее.

В наиболее развитых средствах аннотирования учитывается также зависимость предложений друг от друга с тем, чтобы не вносить в аннотацию обрывки, начинающиеся, например, со слов «К тому же...», «В-третьих...» и тому подобное. Чтобы аннотация получилась связной, программа подбирает группы взаимосвязанных (взаимозависимых) предложений, а затем «склеивает» их, для большей связности немного изменяя на стыках [7].

### 5. Концептуальное аннотирование

В данной работе предлагается подход ТМ к обработке текстовой информации в качестве построения так называемых семантических профилей документов для задач КА.

Онтологический подход к разработке такой системы позволит решать и комплекс прикладных задач, например, поиск экспертов в интересующей области знаний, выявление основных научных приоритетов отдельных ВУЗов, кафедр, преподавателей (создание семантических профилей). Под семантическим профилем документа будем понимать набор отобранных терминов с отсылкой документа к соответствующим понятиям, представленным в онтологической модели ПрО. Механизм КА предполагает связывание полученных терминов с концептами онтологии [8].

Примерами использования семантических профилей документов являются: построение профилей преподавателей с возможностью просмотра, например, основных публикаций, направлений научных исследований; поиск экспертов в заданной ПрО; выявление основных научных приоритетов отдельных ВУЗов, кафедр, преподавателей. Общий метод построения представлен на рис. 1.

Далее рассмотрим все шаги предложенного алгоритма подробнее.

Основной задачей лексического анализа является разбиение входного текста, состоящего из последовательности одиночных символов, на последовательность слов или лексем, то есть выделить эти слова из непрерывной последовательности. Все символы входной последовательности с этой точки зрения разделяются на символы, принадлежащие каким-либо лексемам, и символы, разделяющие лексемы (разделители) [9].

С точки зрения дальнейших фаз анализа лексический анализатор выдает информацию двух сортов: для синтаксического анализатора, работающего вслед за лексическим, существенна информация о последовательности классов лексем, ограничителей и ключевых слов, а для контекстного анализа, работающего вслед за синтаксическим, важна информация о конкретных значениях отдельных лексем (идентификаторов, чисел и так далее). В данной системе на этапе лексического анализа предложено использовать методологию извлечения основных терминов, характеризующих специфическую область исследования.

После того как произведен анализ каждого слова, начинается анализ отдельных предложений (синтаксический анализ), позволяющий определить взаимосвязи между отдельными словами и частями предложения. Результатом такого анализа является граф, узлами которого выступают слова предложения; при этом, если два слова связаны каким-либо образом, то соответствующие им вершины графа связаны дугой с определенной окраской. Возможные окраски дуг зависят от языка, на котором написано предложение, а также от выбранного способа представления синтаксической структуры предложения.

Контекстный, или семантический, анализ текста базируется на результатах синтаксического анализа, получая на входе уже не набор слов, разбитых на предложения, а набор деревьев, отражающих синтаксическую структуру каждого предложения. Поскольку методы синтаксического анализа пока мало изучены, решения целого ряда задач семантической обработки текста базируются на результатах анализа отдельных слов, и вместо синтаксической структуры предложения, анализируются наборы стоящих рядом слов.

Семантический анализ, так или иначе, работает со смыслом слов. Следовательно, должна быть какая-то общая для всех методов анализа база, позволяющая выявлять семантические отношения между словами. Такой основой является тезаурус языка. На математическом уровне он представляет собой ориентированный граф, узлами которого являются слова в их основной словоформе. Дуги задают отношения между словами и могут иметь ряд окрасок.

В качестве одного из методов для кластеризации в данном случае целесообразно использовать k-кластеризацию. Это неиерархический алгоритм, кластеры представлены в виде центроидов, являющихся «центром массы» всех объектов, входящих в кластер, в которой каждый объект связан только с одной группой.

После указанных шагов система получает набор терминов, описывающих основные понятия документа. Под термином в данном случае понимаем слово или словосочетание, точно обозначающее понятие и его соотношение с другими понятиями в пределах исследуемой ПрО. В общем случае термины служат специализирующими, ограничительными обозначениями характерных для ПрО, явлений, их свойств и отношений. В идеале термин должен быть однозначным, систематичным, стилистически нейтральным.

На последнем шаге работы происходит связь концептов онтологии с полученным набором терминов. При этом предполагается наличие первоначальной версии обобщенной онтологии – концептуальной модели исследуемой ПрО, которая в процессе обработки коллекции текстов может редактироваться.

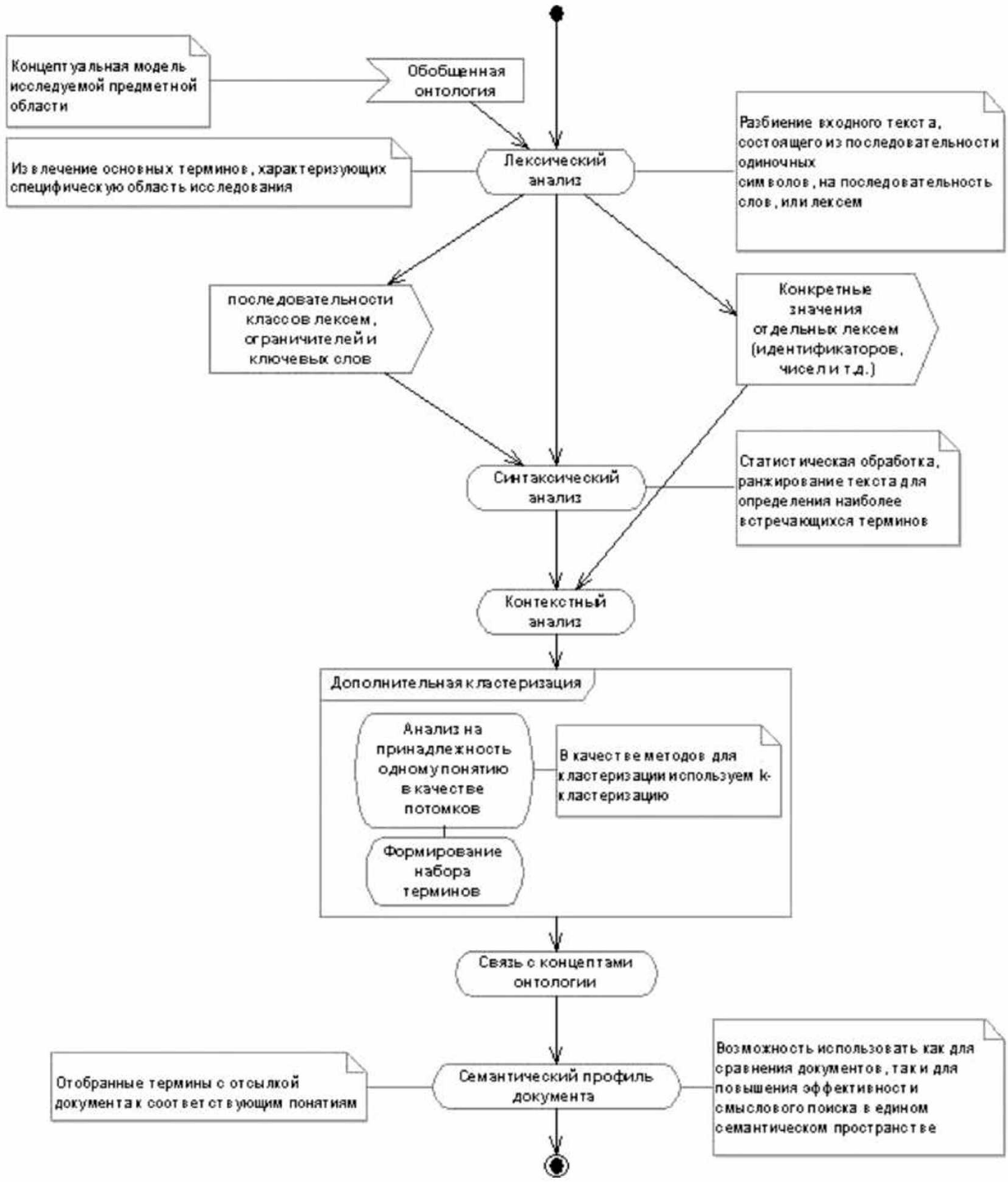


Рис. 1. Общая схема КА

В качестве инструментария создания и редактирования таких онтологий предлагается использование визуальной среды Protégé 3.1 со встроенным плагином PROMPT 2.4.8, который позволяет управлять различными версиями онтологии, осуществляя их слияние и выравнивание [10].

Согласно предложенному алгоритму разработана обобщенная архитектура системы, которая сможет реализовать все описанные функции (рис. 2). Знания системы включают предметные, априорные, фактические (из текстов) и лингвистические. Система включает следующие взаимосвязанные модули: базу фактов, основанную на текстах ПрО, словарь терминов и словарь конечного набора терминов, формируемый в процессе работы системы, а также онтологию, описывающую взаимосвязь понятий в данной сфере.


Рис. 2. Обобщенная архитектура системы КА

**Выводы**

Проанализированы основные задачи, решаемые сегодня системами анализа текстов: формирование информационного портрета текста в терминах ключевых понятий; выявление смысловых связей между понятиями; автоматическое реферирование. Большинство существующих методов ТМ, основу которых составляют словари терминов, используют «плоские» тематические словари – набор ключевых слов, выделенных с помощью статистических методов обработки информации и сгруппированных в иерархическую структуру. Такие словари формируют статистический образ текста.

В работе предложен метод КА как перспективный, мало изученный подход в задачах анализа текстовой информации. Этот метод является основой для создания семантического профиля документа, который предполагает не только статистический анализ текста с определением наиболее часто встречающихся терминов, но и их дополнительную кластеризацию, и анализ на принадлежность одному понятию в качестве потомков.

В качестве понятий выступают концепты онтологии ПрО. Семантические профили в дальнейшем используются для сравнения документов, а также повышения эффективности смыслового поиска в едином семантическом пространстве.

На основе предложенного метода разработан алгоритм КА, подробно рассмотрены основные шаги для построения семантического профиля документа, а также его дальнейшего связывания с концептами онтологии ПрО. Кроме того, в данной работе была разработана архитектура системы КА.

**Список литературы:** **1.** *Гладун А.Я, Рогушина Ю.В.* Онтологии в корпоративных системах // Корпоративные системы: журнал. –Киев: Изд. дом «Комиздат», 2006.–№1. **2.** *Uren V., Cimiano Ph., Iria J., Handschuh S., Vargas-Vera M., Motta E., Ciravegna F.* Semantic annotation for knowledge management: Requirements and a survey of the state of the art // Web Semantics: Science, Services and Agents on the World Wide Web. –2006. – Vol. 4, No. 1. –P. 14-28. **3.** *Mertensson M.* A critical review of knowledge management as a management tool // Journal of Knowledge Management. – 2000. –V. 4, № 3. **4.** *Feldman R., Sanger J.* The text mining handbook: advanced approaches in analyzing unstructured data. –Cambridge University Press, 2007. - 410 p. **5.** *Рябова Н.В., Волошина Н.А., Шубкина О.В.* Применение методов text mining при решении задач менеджмента знаний // Научная сессия МИФИ-2008. Сб. науч. тр. В 15 томах. Т. 10. Интеллектуальные системы и технологи. –М.: МИФИ, 2008. –С.31-32. **6.** *Ильин Н., Киселев С., Рябышкин В., Танков С.* Технологии извлечения знаний из текста // Открытые системы. СУБД. - 2006. –№ 6.–С. 48-53. **7.** *Ашманов И.* Информация и знания: невидимая грань.– http://newasp.omskreg.ru/intellect/f5.htm. **8.** *Шубкина О.В.* Методы и модели построения семантических профилей документов // Материалы 12-ого Междунар. молод. форума «РАДИОЭЛЕКТРОНИКА И МОЛОДЕЖЬ В XXI веке». В 3 частях. Ч.2.–Харьков. –С. 173. **9.** *Компаниец Р.И. и др.* Системное программирование. Основы построения трансляторов. –СПб.: КОРОНА принт, 2000. –256 с. **10.** www.protege.stanford.edu. **11.** *Гаврилова Т.А., Хорошевский В.Ф.* Базы знаний интеллектуальных систем. –СПб: Питер, 2000. –384 с. **12.** *Luger G.* Artificial Intelligence: Structures and Strategies for Complex Problem Solving, 5th edition. –UK: Pearson Education, 2006. –889 p. **13.** *Tiwana A.* The knowledge management toolkit: orchestrating IT, strategy, and knowledge platforms. –2nd ed. –Prentice Hall PTR, 2002. –388 p. **14.** *Feldman R., Sanger J.* The text mining handbook: advanced approaches in analyzing unstructured data. –Cambridge University Press, 2007. –410 p.

***Поступила в редколлегию 31.03.2009***

УДК 519.52

**Інтелектуальний аналіз текстів в системах менеджменту знань** / О. В. Шубкіна // Біоніка інтелекту: наук.-техн. журнал.–2009.–№1(70).–С. 142–146.

Розглядаються основні задачі систем аналізу текстової інформації. Запропонований метод концептуального анотування, визначені його головні етапи. Розглянута узагальнена архітектура системи концептуального анотування.

Іл.: 2. Бібліогр.: 14 найм.

UDK 519.52

**Text mining in the knowledge management systems** / O. V. Shubkina // Bionics of Intelligence: Sci. Mag. –2009. –№1(70).–P. 142–146.

The main aims of the text mining systems are considered. Method of the concept annotation is offered, its basic stages are defined. Generic architecture of the concept annotation system is considered.

Fig.: 2. Ref.: 14 items.