электронное научно-техническое издание
НАУКА и ОБРАЗОВАНИЕ
Эл № ФС 77 - 30569. Государственная регистрация №0421100025. ISSN 1994-0408

Задачи управления знаниями, извлекаемыми из текстовых документов.

09, сентябрь 2011

автор: Норенков И. П.

УДК 519.6

МГТУ им. Н.Э. Баумана
norenkov@wwwcdl.bmstu.ru

Введение. Под управлением знаниями (KM - knowledge management) понимают процессы сбора, хранения, извлечения и обработки знаний в целях совершенствования деятельности предприятий и организаций [1, 2]. Значительная часть полезных для организации знаний содержится в документальных базах знаний (БЗ), являющихся базами текстовых документов. Поскольку число документов в БЗ большинства организаций весьма значительно и продолжает расти, ручное управление знаниями оказывается неэффективным. Поэтому в интеллектуальных системах управления знаниями стремятся в максимально возможной степени автоматизировать процедуры KM. Однако текстовые документы и соответственно знания, содержащиеся в них, обычно являются слабо структурированными, что существенно сокращает возможности автоматических извлечения и обработки знаний в системах KM.

Обработка знаний может быть направлена на достижение ряда целей, что порождает ряд задач обработки знаний, решаемых в интеллектуальных системах. К ним, в первую очередь, относятся задачи кластеризации, классификации, аннотирования, упорядочеиия документов, поддержки принятия решений. В силу слабой структурированности знаний успешно решать эти задачи в автоматическом или полуавтоматическом режимах удается лишь в отдельных частных случаях, поэтому существующие системы KM, как правило, являются узкоспециализированными.

Повышению эффективности решения интеллектуальных задач способствует применение онтологий. Модели знаний в онтологиях выражены в виде множеств понятий (концептов, сущностей) и отношений между ними. Поскольку извлечение знаний из текстовых документов и их обработка связаны именно с понятийным составом рассматриваемых приложений, управление документальными БЗ целесообразно выполнять на основе применения онтологий.

В статье приведен обзор основных подходов к решению задач извлечения и обработки знаний в интеллектуальных системах. В их числе рассмотрен предложенный метод, основанный на кластеризации онтологий.

Классификация и кластеризация документов. Очевидный метод классификации документов предполагает предварительную ручную разработку пользователями-экспертами обучающей выборки, состоящей из специально отобранных документов [3]. Далее выполняется формирование некоторых классификационных признаков и обучение классификатора. На основе выявленных классификационных признаков формируются правила оценки тематической направленности документа, используемые для классификации приходящих в систему документов [4].

Преимущества онтологического подхода заключаются в том, что вместо разработки обучающей выборки создаются прикладные онтологии, которые могут использоваться не только для классификации, но и для решения ряда других задач. В системах КМ подразумевается предварительная разработка предметных (domain) онтологий, охватывающих области деятельности организации. Если перечни приложений при предметной кластеризации онтологий и документов совпадают, то наличие предметных онтологий позволяет легко осуществить автоматическое распределение документов по рубрикам, т.е. автоматическую кластеризацию документов.

При классификации оценка R_{kl} степени принадлежности k-го документа l-му тематическому кластеру выполняется по n_{kl} – числу появлений концептов l-й предметной онтологии в k-м документе; $k = 1, 2,\ldots q$; $l = 1, 2,\ldots m$; например:

$$R_{kl} = \frac{n_{kl}}{\sum_{i=1}^{m} n_{ki}},$$

где $n_{kl} = \sum_{j=1}^{M_j} N_{jkl}$; N_{jkl} - число появлений j-го концепта l-й предметной онтологии в k-м документе, M_j — число концептов в l–м приложении.

Эта оценка может использоваться для классификации документов с введением весов концептов, учитывающих их информативность. Так, в работе [5] классификация осуществляется по суммарному весу концептов предметной онтологии, имеющихся в документе. Вес g_j j-го концепта часто определяют следующим образом:

$$g_j = \ln((q+1)/(q_j+1)),$$

где q_j – число документов в базе, содержащих j-й дескриптор (термин, обозначающий концепт).

Аннотирование документов. Аннотирование является основной задачей извлечения информации (IE - Information Extraction) из текстовых документов. Аннотирование, которое можно трактовать как составление метаданных документов, является основой для решения ряда задач обработки знаний. Подходы к аннотированию документов произвольной тематики и узкой направленности различны.

Наиболее сложны задачи автоматического извлечения информации из документов неструктурированных или слабо структурированных, причем трудности IE возрастают по мере расширения тематики исследуемого корпуса документов. В существующих системах попытки извлечения знаний из таких документов основаны на выявлении в текстах паттернов (словосочетаний, предложений), содержащих определенное ключевое слово (обычно глагол) вместе с сопутствующими словами, выполняющими такие роли, как «субъект», «инструмент», «цель» [6]. Одним из условий применения этого подхода для автоматического аннотирования текстов произвольной тематики является их синтаксическая и семантическая корректность.

Чаще аннотирование выполняется для структурированных документов конкретной тематики. В первой группе методов, ориентированных на такое аннотирование, используется поиск и выявление в документах часто встречающихся слов, характеризующих конкретные события, ситуации, факты. К таким словам относятся экземпляры концептов, такие как собственные имена (NE - named entities), названия организаций, географических пунктов, даты, адреса и т.п. Во второй группе методов извлечение информации заключается в поиске специфических выражений, характерных для определенных предметных областей [7]. Полуавтоматическое аннотирование документов и извлечение нужных данных при этом обычно происходит на основе предварительного обучения системы IE, выполняемого пользователем [8].

Аннотирование с применением онтологий позволяет составлять аннотацию из терминов концептов или значений (экземпляров) концептов, найденных в тексте документа [9, 10]. Здесь по-прежнему популярно аннотирование на основе использования NE. Аннотация представляет собой сформированные высказывания, содержащие NE и выраженные в формате RDF [11]. Аннотации в форме онтологий, что позволяет использовать средства онтологического анализа как для самих документов, так и для аннотаций, рассматриваются в работе [12].

Информационный поиск. Информационный поиск (IR - Information Retrieval) лежит в основе решения большинства задач управления знаниями. Методы и средства IR освещены в большом числе работ, например, в монографии [13].

В настоящее время преимущественно используется векторная модель информационного поиска [14], основанная на сопоставлении поисковых образов запроса ПОЗ и документа ПОД:

$$\text{ПОД} = \{x_1, x_2, \ldots x_{n1}\},$$

$$\text{ПОЗ} = \{y_1, y_2, \ldots y_{n2}\},$$

где x_i € $\mathbf{X}$ – термин (слово или словосочетание) в тексте документа; y_i € $\mathbf{X}$ – термин в поисковом запросе, $\mathbf{X}$ – множество слов из используемого словаря поисковой системы, за вычетом стоп-слов; $n1$ и $n2$ – числа ключевых слов, вошедших соответственно в ПОД и ПОЗ. Индекс поисковой системы состоит из списка слов множества $\mathbf{X}$, каждому элементу x_j списка соответствует множество ссылок на документы, в которых присутствует x_j, $j = 1, 2, \ldots |\mathbf{X}|$. Релевантность запроса и k-го документа либо определяется по формуле

$$r_k = \frac{\sum_{i=1}^{h}(g_i\ z_{ki})}{\sum_{i=1}^{h} g_i},$$

где g_i – вес i–го ключевого слова x_i запроса, $z_{ki} = 1$ при наличии x_i в тексте документа, иначе $z_{ki} = 0$, h – число слов в запросе, либо по косинусу угла между частотными векторами запроса и документа. В качестве веса термина часто используют параметр TF-IDF, равный отношению частоты упоминания слова в данном документе к частоте употребления этого слова в остальных документах коллекции [13].

Онтологии применяются для повышения эффективности поиска. Например, в [15] концепты онтологии, имеющиеся в ПОЗ и ПОД, представляются в виде вершин поддеревьев и релевантность определяется сопоставлением этих поддеревьев. Также на основе анализа графовых моделей релевантность запроса и документа определяется в работе [16]. В [17] предлагается при поиске в Интернет использовать персонифицированные онтологии, формируемые на основании анализа поведения пользователя в процессе IR.

В работе [18] излагается подход, основанный на совместном применении ГРНТИ и методов онтологического моделирования. Тематические кластеры соответствуют градациям ГРНТИ. В кластерах онтологий учитываются отношения род-вид, часть-целое и синонимии. *Аннотация* документа составляется из его названия и ключевых слов. *Классификация* выполняется по аннотации - определяется принадлежность документа определенному кластеру (позиции ГРНТИ). Затем на этапе индексации в *индекс* включаются не все значащие слова, а только концепты соответствующей предметной онтологии, обнаруживаемые в тексте документа. Запрос формируется из предъявляемых пользователю концептов выбранной им рубрики (кластера).

Упорядочение документов. Упорядочение документов возможно по тем или иным показателям. Так, в индексах систем информационного поиска ссылки на документы обычно упорядочены по степени их важности, например, с использованием показателя TF-IDF. Часто требуется упорядочение по таким параметрам, как дата написания документа, фамилия автора (алфавитный порядок) и т.п.

Важной для успешной деятельности организаций и предприятий задачей является повышение квалификации персонала и, следовательно, отбор и упорядочение используемых для этих целей учебных материалов. Формирование электронных учебных пособий для обучающих систем на основе применения онтологий рассмотрено в [19]. Технология отбора и семантического упорядочения учебных текстов реализована в системе БиГОР, в которой документы, как разделяемые учебные модули, являются интерпретаторами концептов [20].

Поддержка принятия решений. Основной целью управления знаниями является поддержка принятия решений. Системы поддержки принятия решений (СППР) могут быть основаны на правилах (RBR – Rule-Based

Reasoning) [21] или на прецедентах (CBR - Case-Based Reasoning) [22]. Используется также комбинация этих двух подходов [23].

Создание СППР подразумевает разработку базы знаний и системы инструментальных средств, реализующих построение БЗ, визуализацию, структурирование информации кластеризацию, упорядочение знаний, информационный поиск, а при применении онтологического подхода также построение онтологий [24].

Системы RBR преимущественно реализуются в виде экспертных систем. Обычно правила принятия решений удается сформулировать только для отдельных конкретных приложений, поэтому системы RBR являются узкоспециализированными системами.

В системах CBR для принятия решения используется ранее накопленный опыт принятия решений в аналогичных ситуациях, выраженный в виде описаний прецедентов. Прецеденты-решения могут быть представлены в той или иной форме, принятой в интеллектуальных системах, например, в виде фреймов, семантических сетей, значений параметров и т.п. [25, 26]. База прецедентов в СППР на основе CBR состоит из пар «проблема/решение», причем решение чаще всего представлено фреймом, в слотах которого содержатся значения параметров, характеризующих условия и результаты решения задачи, в частности, результатами могут быть значения предметных переменных. Задача формирования фрейма становится задачей извлечения знаний из тех или иных источников. При извлечении знаний из текстовых документов, а также при описании решений и при сравнительной оценке прецедентов полезно используются онтологии.

Автоматическое построение базы прецедентов возможно только для конкретных структурированных приложений. Существуют прагматический и содержательный подходы к управлению знаниями [27]. Прагматический подход характерен для обработки структурированной информации с использованием баз данных. Содержательный подход имеет место при использовании специальных моделей и операций представления и обработки знаний.

При работе с неструктурированной информацией, относящейся к сравнительно широким предметным областям, неизвестна сама структура фрейма «проблема/решение», возможности автоматической или полуавтоматической обработки знаний резко сужаются. На основе содержательного подхода удается формализовать фазу извлечения знаний, фаза собственно принятия решений остается неавтоматизированной.

Ролевая кластеризация онтологий. Описываемый далее метод ролевой кластеризации онтологий можно отнести к числу реализаций содержательного подхода. Метод основан на предварительном распределении концептов онтологии по кластерам в зависимости от их роли в сложных концептах.

Сложные концепты – ‘это словосочетания, выражающие отношения «объект-свойство», «объект-действие», «объект-свойство-действие», «средство-действие-объект» и состоящие из простых концептов, выполняющих роли «объект», «свойство», «действие», «средство» [19]. Хотя распределение концептов по кластерам не является формальной процедурой, но выполняется для каждого приложения однократно. Сложные концепты представляются в виде паттернов, слоты которых соответствуют ролям простых концептов.

Метод ролевой кластеризации онтологий перспективен для полуавтоматического решения задач управления знаниями, поскольку сложные концепты гораздо более точно выражают семантику документов и запросов, чем совокупности составляющих их простых концептов.

Аннотации документов в соответствии с методом ролевой кластеризации составляются из сложных концептов или из предложений, в состав которых входят сложные концепты. Извлечение сложных концептов из текста заключается в поиске терминов, относящихся к разным кластерам онтологии и расположенных в нужной последовательности близко друг от друга в тексте документа, например, в пределах одного предложения. Найденные предложения со сложными концептами включаются в аннотацию (с возможной ручной корректировкой), если частота повторения концептов в документе не ниже заданного порога. Возможно совместное применение аннотирования на основе сложных концептов и на основе специфических терминов типа NE.

Информационный поиск с использованием ролевой кластеризации концептов выполняется по запросам, содержащим сложные концепты. Сложные концепты выделяются в запросе (или запрос конструируется по паттернам сложных концептов), далее запрос сопоставляется с аннотацией документа. Другими словами, вместо поиска по ключевым словам (простым концептам) выполняется поиск по сложным концептам, входящим в индекс, а для индексирования используются результаты аннотирования.

Для принятия решений с использованием ролевой кластеризации онтологий применяется метод CBR, в соответствии с которым каждому документу коллекции сопоставлен паттерн проектирования «проблема/решение». Метаданные документа, включая аннотацию, составляют автоматически формируемую левую часть паттерна, т.е. слоты «проблема». Правая часть в случае структурированных приложений содержит значения параметров прецедентов. В случае неструктурированных приложений метод обеспечивает поиск документов, относящихся к проблеме, выраженной сложными концептами, и потенциально содержащих прецеденты решения проблемы. Окончательное выявление прецедентов возлагается на пользователя.

Заключение. Большинство задач извлечения знаний из текстовых документов и их обработки решается в настоящее время вручную с частичным применением полуавтоматических методов анализа информации и принятия решений. Актуальность разработки автоматизированных методов управления знаниями повышается в связи с ростом объема документальных баз знаний, трудностями поиска в них полезной информации и экономической целесообразностью многократного использования ранее разработанных эффективных и описанных в документах методов и средств решения сложных задач. С помощью метода, основанного на кластеризации онтологий, удается автоматизировать, во-первых, фазу извлечения информации из документов и, во-вторых, определение фрагментов текста, перспективных для первоочередного ручного анализа на предмет выявления описаний проектных решений.

Литература

1. Dieng, R., Corby, O., Giboin, A., & Ribière, M. Methods and Tools for Corporate Knowledge Management. // In International Journal of Human-Computer Studies, 1999, 51 (3), pp. 567-598.

2. Гаврилова Т. Извлечение знаний: лингвистический аспект // Enterprise Partner, 2001 г. №10.

3. Шабанов В.И., Андреев А.М., Метод классификации текстовых документов, основанный на полнотекстовом поиске // Труды РОМИП’2003ю - СПб: НИИ Химии СПбГУ, 2003,- с. 52-71.

4. Oracle Text Application Developer's Guide. - Oracle Corporation, 2003. http://www.stanford.edu/dept/itss/docs/oracle/10g/text.101/b10729.pdf

5. Nagarajan M., Sheth A., Aguilera M., Keeton K., Merchant A., Uysal M.. Altering Document Term Vectors for Classification - Ontologies as Expectations of Co-occurrence // 16th World Wide Web Conference, 2007, pp 1225-1226.

6. Muslea I. Extraction Patterns for Information Extraction Tasks: A Survey // In AAAI-99 Workshop on Machine Learning for Information Extraction, 1999.

7. Li Y., Zhang L., Yu Y. Learning to Generate Semantic Annotation for Domain Specific Sentences. // In: K-CAP 2001 Workshop on Knowledge Markup & Semantic Annotation, 2001.

8. Kushmerick N., Weld D., Doorenbos R.. Wrapper Induction for Information Extraction. // In Proc. 15th Int. Joint Conf. AI, 1997, pp 729-735.

9. Fernández M., Vallet D., Castells P. Automatic Annotation and Semantic Search from Protégé // In: 8th International Protege Conference, 2005, Madrid, Spain.

10. Handschuh1 S., Staab1 S.,, Ciravegna F.. S-CREAM - Semi-automatic CREAtion of Metadata // In Proc. of the European Conference on Knowledge Acquisition and anagement EKAW-2002. Madrid: Springer, 2002.

11. Kiryakov, A., Popov, B., Terziev, I., Manov, D., Ognyanoff, D. Semantic Annotation, Indexing, and Retrieval // Journal of Web Sematics, Issue 1, 2005, pp 47-49.

12. Castells P., Fernandez M., Vallet D. An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval. // IEEE Transactions on Knowledge and Data Engineering 19(2), Special Issue on Knowledge and Data Engineering in the Semantic Web Era, February 2007, pp. 261-272.

13. Manning C., Raghavan P., Schütze H. Introduction to Information Retrieval. - Cambridge University Press. 2008. 544 p.

14. Salmon G., McGill M. Introduction to Modern Information Retrieval. - McGraw Hill, New York, 1986.

15. Baziz M., Boughanem M., Pasi G., Prade H., An Information Retrieval Driven by Ontology from Query to Document Expansion // Conference RIAO2007, Pittsburgh PA, U.S.A. 2007.

16. Карпенко А. П. Оценка релевантности документов нтологической базы знаний // Электронное научно-техническое издание «Наука и образование», 2010, № 9.

17**.** Calegari S., Pasi G. Ontology-Based Information Behaviour to ImproveWeb Search // Future Internet, 2010, *2*, pp 533-558.

18. Вдовицын В.Т., Лебедев В.А.. Технологии систематизации и поиска электронной научной информации с применением онтологий // Информационные Ресурсы России, 2010, № 5

19. Норенков И.П. Документальные базы знаний на основе онтологий // Информационные технологии, 2011, № 2, с. 11-16

20. Норенков И.П., Уваров М.Ю. База и генератор образовательных ресурсов // Информационные технологии, 2005, № 9, с. 60-65.

21. Ligeza A. Logical Foundations for Rule-Based Systems. // Series "Studies in Computational Intelligence", v. 11. - Springer-Verlag, 2006, 329 p.

22. Aamodt A., Plaza E. Case-Based Reasoning: Foundational Issues, Methodological Variations and System Approaches" // AI Commutations, IOS Press, 1994, v. 7, 1, pp. 39-59.

23. Prentzas J., Hatzilygeroudis I. Categorizing Approaches Combining Rule-Based and Case-Based Reasoning'// Expert Systems, 2007, № 24, pp 97-122.

24. Ситников П.В. Построение систем поддержки принятия решений на основе онтологий // Автореферат диссертации на соискание ученой степени кандидата технических наук. – Самара: 2009, 24 с.

25. Watson I., Marir F.Case-Based Reasoning: A Review // Knowledge Engineering Review, V. 9, No. 4, 1994, pp 355-381.

26. Kurbalija V., Budimac Z. Case-Based Reasoning Framework for Generating Decision Support Systems. // Novi Sad J. Math., V. 38, No. 3, 2008, pp 219-226.

27. Ермаков А.Е. Извлечение знаний из текста и их обработка: состояние и перспективы // Информационные технологии, 2009, № 7, с. 50-55.