# НЕКОТОРЫЕ ПЕРСПЕКТИВЫ ИСПОЛЬЗОВАНИЯ ЭЛЕКТРОННЫХ БИБЛИОТЕК В КАЧЕСТВЕ ИСТОЧНИКОВ ЗНАНИЙ ДЛЯ ЭКСПЕРТНЫХ СИСТЕМ

Смирнов В.В., фирма ОПТИМА, Б.Почтовая, д.26В
к.мед.н. Беляев М.В., фирма «Радение», Ленинградский проспект, д.15
Vitaly_Smirnov@mail.ru

В работе рассматриваются некоторые перспективы использования электронных библиотек в качестве источников знаний для экспертных систем. Рассматриваются возможности и ограничения для применения методов извлечения знаний из текстов и Data Mining к электронным библиотекам. Предлагается подход к извлечению знаний из электронных библиотек, удовлетворяющий указанным ограничением.

# SOME PERSPECTIVES FOR DIDGITAL LIBRARIES USAGE AS KNOWLEDGE SOURCES FOR EXPERT SYSTEMS

Smirnov V.V., OPTIMA, Bolshaya Pochtovaya, 26B
Beljaev M.V., RADENIE, Leningradskay prospect, 15
Vitaly_Smirnov@mail.ru

Some perspectives for digital libraries usage as knowledge sources for expert systems are described in this paper. The application abilities and restrictions of methods for knowledge elicitation from texts and Data Mining in digital libraries are examined. It is proposed the approach to knowledge elicitation from digital libraries according to mentioned restrictions.

## Введение

Процесс создания экспертной системы (ЭС) включает в себя извлечение знаний из источников знаний, в качестве которых могут выступать эксперты, тексты, базы данных (БД) и некоторые другие. Электронные библиотеки (ЭБ) с одной стороны, могут содержать тексты, а с другой – представлять собой БД, что означает, что они могут одновременно обладать свойствами источников двух типов. Для каждого из этих типов источников разработана своя группа методов извлечения знаний. Поэтому рассмотрим перспективы независимого применения каждой группы методов к ЭБ, а также их совместного использования.

**Сходство и отличие методов извлечения знаний из текстов и методов обработки текстов в ЭБ**

Существует сходство методов извлечения знаний из текстов и обработки текстов в ЭБ, рассмотрим некоторые примеры. Методы, основанные на представлении текстов в виде семантических сетей с одной стороны, используются в ЭБ для кластеризации документов [Когаловский, 1999], с другой стороны, используются для формирования базы знаний (БЗ) [Осипов, 1997]. Аннотирование является одним из видов обработки документов ЭБ. В редакторе протоколов интегрированной среды KADS [Schreiber et al., 1988; Schreiber et al, 1993] для поддержки проектирования ЭС также существует возможность связать фрагмент протокола интервьюирования эксперта или протокола «мыслей вслух» с некоторой аннотацией. В системе KRITON [Diederich et al., 1987] поддерживается возможность запросить статистическую информацию о частотных характеристиках ключевых слов в тексте и определение размера фрагмента, окружающего ключевое слово, что сходно с анализом содержания текстов в системе TextAnalyst [TextAnalyst], одной из функций которой является автоматическое выделение основных понятий произвольных текстов (словосочетаний и слов), а также их взаимосвязей с оценкой относительной значимости.

В качестве отличий методов извлечения знаний и обработки текстовых документов в ЭБ можно выделить следующие:

- процесс извлечения знаний из текстов, в отличие от обработки текстов в ЭБ, часто выполняется с участием эксперта (например, в рамках технологии, изложенной в работе [Осипов, 1997];
- в качестве методов обработки текстов при извлечении знаний используются, как правило, методы лексико-семантического анализа, а также модели понимания текста на лингвистическом и семантическом уровнях [Осипов, 1997]; в ЭБ наблюдается большее разнообразие, т.к. наряду с семантической обработкой текстов используются статистические методы, методы поиска и др;
- в ЭБ обработку проходят все тексты, которые можно разделить на художественные и специальные [Валькман и др., 2000]; для извлечения знаний используются специальные тексты, в частности, носители профессиональных знаний (учебники, методические материалы, статьи, монографии, инструкции и т.п.) [Гаврилова и др., 1992];
- результаты обработки текстов ЭБ могут помещаются либо в БД, либо в БЗ, см. например, [Валькман и др., 2000], а результаты извлечения знаний из текстов обычно помещаются в БЗ, например, [Gomez et al., 1990].

Таким образом, можно найти ЭБ, для которой обработка текстов сходна с извлечением из них знаний, поэтому в перспективе использование обработанных в ЭБ текстов может упростить применение к ним методов извлечения знаний.

Известно, что методы извлечения знаний из текстов наименее разработаны [Осипов, 1997]. С другой стороны, в ЭБ используется более широкий спектр методов обработки текстов, охватывающих как художественные, так и специальные тексты. Поэтому можно говорить о перспективе интеграции ЭБ со средствами извлечения знаний из текстов с целью развития последних.

**Перспективы использования технологии Data Mining в ЭБ**

Data Mining (интеллектуальный анализ данных) используется для решения различных задач построении ЭБ, например, [Барышев и др., 2001;Вдовицын и др., 2001]. В настоящее время технология Data Mining часто используется для формирования БЗ ЭС. Например, [Загоровский и др., 2000; Рыбина и др., 2000]. Описание задач, для которых применяется DM можно найти в работах [Забежайло, 1998; Загоруйко, 2000]. В качестве примеров алгоритмов Data Mining, используемыми для построения БЗ, можно назвать A1 [Breiman et al., 1994], CN2 [Clarck et al., 1988], ID3 [Quinlan, 1986], OC1 [Murthy et al., 1994], C4.5 (расширение ID3) [Quinlan, 1992], ДРЕВ [Вагин, 1988], Ripper [Cohen, 1995]. Сравнение алгоритмов DM можно найти в работах [Куликов и др., 2000; Рыбина и др., 2000; Nahm, 2000].

Существует ряд трудностей использования алгоритмов Data Mining в ЭБ, в частности:

- наличие в ЭБ большого количества неструктурированных текстов;
- наличие скрытых связей между содержимым текстов и данными, сопровождающими эти тексты (набора атрибутов, графических изображений, и др.);
- наличие скрытых связей между текстами.

Одним из подходов к преодолению этих трудностей может быть интеграция методов DM с методами обработки текстов. Существуют технологии, в которых предусмотрена интеграция извлечения знаний из текстов с DM. В частности, такая интеграция выполнена в рамках технологии построения систем, основанных на знаниях, изложенной в работе [Осипов, 1997], которая объединяет систему прямого приобретения знаний от экспертов SIMER [Осипов, 2001], программу моделирования рассуждений MIR, программу выявления семантических связей из текста EXTRA и программу выявления знаний из данных KAD. Однако, в соответствии с этой технологий, объединение результатов обработки текстов и анализа данных выполняется при участии эксперта, что создает трудности при обработке большого количества текстов. Другим примером интеграции методов извлечения знаний из текстов с методами DM является задачно-ориентированная методология автоматизированного построения экспертных систем для статических проблемных областей [Рыбина, 1997; Рыбина,

1998; Рыбина, и др., 2000]. Сложность ее применения для извлечения знаний из ЭБ состоит в том, что существующие в настоящее время версии поддерживающего эту методологию комплекса АТ-ТЕХНОЛОГИЯ [Рыбина и др., 1997; Рыбина, 1998] при извлечении знаний из БД (на основе алгоритмов CN2, OC1 или ID3 [Рыбина, и др., 2000]) не учитывают возможность присутствия в ней текстовых документов.

Другим полезным подходом может являться подход на основе Text Mining (TM), то есть процесса поиска полезных или интересных примеров, моделей, директив, тенденций или правил из неструктурированного текста. В качестве примеров работ в области TM можно привести [Feldman et al., 1995; Tanabe et al., 1999].

Интересным представляется подход в рамках направления TM, в соответствии с которым извлекаемая из текстов информация помещается в БД, и к полученной БД затем применяются методы DM. При этом, БД может быть сформирована методами Information Extraction (IE), основной задачей которых является извлечение существенных фактов о предварительно заданных типах событий, сущностях или отношениях. Примерами систем IE могут служить FLORID [Wolfgang, 2000] и FASTUS [Hobbs et al., 1992]. В работе [Nahm et al., 2000] предложен подход, в соответствии с которым сформированная в результате IE БД используется для формирования БЗ прогнозирующих правил методами DM (C4.5 или Ripper).

Подход, предложенный в работе [Nahm et al., 2000], взят за основу при наполнении БЗ разрабатываемой в настоящее время экспертно-обучающей системы «Эндоскопист», предназначенной для обучения врачей эзофагогастродуоденоскопии и колоноскопии. Одним из компонентов системы «Эндоскопист» является коллекция типовых изображений нормальной слизистой и нормального взаимоотношения внутренних структур и часто встречающихся патологических процессов в пищеводе, желудке, двендцатиперстной кишке, толстой кишке и терминальном отделе подвздошной. Вместе с изображениями в коллекции хранятся протоколы, содержащие текстовые описания соответствующих изображений. Коллекция изображений и протоколов использована в системе для формирования обучающих воздействий. Кроме того, протоколы применены в качестве источника знаний при формировании БЗ по диагностике патологий.

Процесс извлечения знаний из протоколов, помещенных в коллекцию состоит из следующих этапов:

- автоматизированное извлечение лексики из тестовых примеров протоколов для наполнения словарей лингвистического процессора (ЛП);
- извлечение множества атрибутов, характеризующих объекты эзофагогастродуоденоскопии и колоноскопии, и множества значений атрибутов, используя ЛП, для занесения их в БД;
- формирование фрагментов поля знаний в виде деревьев решений одним из методов DM и предоставление их эксперту для отбора и корректировки;

- добавления фрагментов БЗ в виде наборов продукций, формируемых по отобранным фрагментам поля знаний.

Автоматизированное извлечение лексики реализовано аналогично извлечению лексики в системе НЕВОД [Смирнов, 2001; Смирнов, 2002], которое состоит из следующих основных этапов [Смирнов, 2002]:

- ручное заполнение словарей ЛП базовой лексикой, при котором вводятся семантические классы и категории, заполняются таблицы квазифлексий, вводятся распространенные предикаты с их моделями управления;
- автоматизированное заполнения словарей, при котором из текстов извлекаются понятия и характеристики;
- автоматическое извлечение лексики документов.

Для автоматического извлечения лексики документов в системе НЕВОД применяется набор запросов (фильтров для заданной части предикатно-аргументных структур), при построении которых используется специальный мастер, обеспечивающий создание и редактирование запросов в формате XML с сохранением их в БД, а также позволяющий привести слова-образцы к канонической форме и занести новые слова в словарь. Для системы «Эндоскопист» мастер построения запросов не требуется, так как получаемые в результате обработки текстов предикатно-аргументные структуры целиком заносятся в БД.

В настоящее время проводятся исследования с целью выбора наиболее эффективного алгоритма DM для решения задачи извлечения знаний из данных, полученных путем обработки протоколов.

## Заключение

Анализ методов и средств извлечения знаний из текстов и БД показывает, что существуют перспективные решения, которые могут быть использованы для извлечения знаний из ЭБ, как БД и источника специальных текстов. Причем, результаты обработки текстовых документов в ЭБ могут повысить эффективность извлечения из них знаний. Кроме того, проблема извлечения знаний из БД ЭБ, связанная с наличием в БД ЭБ текстовых документов, может быть решена на основе Text Mining.

## Литература

**[Барышев и др., 2001]** Барышев Д., Высоцкий С. Кукс, Михайлова Е, Некрестьянов И., Новиков Б., Павлова Е. Интеграция публично доступных архивов списков рассылки. // В кн.: Электронные библиотеки: перспективные методы и технологии, элекстронные коллекции. Сборник докладов Третьей Всероссийской конференции. RCDL’2001, Петрозаводск, 11-13 сентября 2001 г. –Карельский научный центр РАН, 2001. С.57-63.

**[Вагин, 1988]** Вагин В.Н. Дедукция и обобщение в системах принятия решений. М.:Наука, 1988.
**[Валькман и др. 2000]** Валькман Ю. Р., Валькман Р. Ю., Золотаревский И. А., Квачев В. Г., Книга Ю. Н., Лозовский В. В., Яковенко Л. П. Международный семинар Диалог'2000 по компьютерной лингвистике и ее приложениям, 2000. Сборник трудов в 2-х томах. Т.2 Прикладные проблемы. http://www.dialog-21.ru/Archive/2000/Dialogue%202000-2/66.htm
**[Вдовицын и др., 2001]** Вдовицын В.Т., Керт Г.М., Беляева Н.А., Луговая Н.Б., Сорокин А.Д., Чуйко Ю.В. Электронная коллекция информационных ресурсов по топонимии европейского севера России. // В кн.: Электронные библиотеки: перспективные методы и технологии, элекстронные коллекции. Сборник докладов Третьей Всероссийской конференции. RCDL'2001, Петрозаводск, 11-13 сентября 2001 г. –Карельский научный центр РАН, 2001.С. 199-201.
**[Гаврилова и др., 1992]** Гаврилова Т.А., Червинская К.Р. Извлечение и структурирование знаний для экспертных систем. М.: Радио и связь.1992.
**[Загоровский и др., 2000]** Загоровский И.М., и др. Королев А.И., Сазонова Л.И., Использование технологии Data Mining для создания системы прогнозирования лавинной опасности. // В кн. КИИ-2000. Седьмая нац. конференция с межд. участием. Тр. конф. М.: Физматлит, 2000, Т.1. С. 91-96.
**[Загоруйко, 2000]** Загоруйко Н.Г. Применение методов Data Mining для ориентации разумного агента в социальной среде многоагентской системы. систем // В кн. КИИ-2000. Седьмая нац. конференция с межд. участием. Тр. конф. М.: Физматлит, 2000, Т.1. С. 97-103.
**[Когаловский, 1999]** Когаловский М.Р. Научные коллекции информационных ресурсов в электронных библиотеках. Труды Первой Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции", С.-Пб. университет, 1999. http://www.cemi.rssi.ru/mei/articles/dlib.htm
**[Куликов и др., 2000]** Куликов А.В., Фомина М.В. Разработка алгоритма обобщения знаний. // В кн. КИИ-2000. Седьмая нац. конференция с межд. участием. Тр. конф. М.: Физматлит, 2000, Т.1. С. 135-142.
**[Осипов, 1997]** Осипов Г.С. Приобретение знаний интеллектуальными системами: Основы теории и технологии. М.:Наука. Физматлит, 1997.
**[Осипов, 2001]** Осипов Г.С. Технология построения распределенных интеллектуальных систем SIMER+MIR. Исследовательский центр искусственного интеллекта ИПС РАН. Фонд эффективной политики. Научно-практическая конференция "Проблемы обработки больших массивов неструктурированных текстовых документов", 2001. http://www.fep.ru/text/dataarrays09.html
**[Рыбина, 1997]** Рыбина Г.В. Задачно-ориентированная методология автоматизированного построения интегрированных экспертных систем для

статических проблемных областей// Изв. РАН. Теория и системы управления. № 5, 1997. С. 129-137.
**[Рыбина и др., 1997]** Рыбина Г.В. Пышагин С.В., Смирнов В.В., Чабаев А.В. Программные средства и технология автоматизированного построения интегрированных экспертных систем. // Программные продукты и системы (Software & Systems). 1997, № 4.
**[Рыбина, 1998]** Рыбина Г.В. Автоматизированное построение баз знаний для интегрированных экспертных систем // Изв. РАН. Теория и системы управления. № 5, 1998. С.152-166.
**[Рыбина и др., 2000]** Рыбина Г.В., Калинина Е.А. Применение технологии Data Mining для автоматизированного построения баз знаний интегрированных экспертных систем // В кн. КИИ-2000. Седьмая нац. конференция с межд. участием. Тр. конф. М.: Физматлит, 2000, Т.1. С. 119-127.
**[Смирнов, 2001]** Смирнов.В.В. Повторное использование словарей при создании новых. // В кн. Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Сборник трудов Третьей Всероссийской конференции по Электронным Библиотекам. RCDL'2001, Петрозаводск, 11-13 сентября 2001 г. Карельский научный центр РАН, 2001, с.73-77.
**[Смирнов, 2002]** Смирнов.В.В. Об одном подходе к извлечению лексики из текстов на ограниченном естественном языке // В кн. Радиоэлектроника, электротехника и энергетика. Восьмая Междунар. науч.-техн. конф. студентов и аспирантов: Тез. докл. В 3-х т. Т.1. М.: Издательство МЭИ, 2002., с.288.
**[Breiman et al., 1994]** Breiman, Friedman, Olshen, Stone. Classifiction and Decision Trees. Wadsworth, 1994.
**[Clarck et al., 1988]** Clark P., Niblett T., The CN2 induction algorithm // Machine Learning Journal. N 4, 1988.
**[Cohen, 1995]** Cohen W.W., Fast effective rule induction. In Proceedings of the Twelfth International Conference on Machine Learning, pp. 115-123, 1995.
**[Diederich et al., 1987]** Diederich, J., Ruhmann, I. and May, M. (1987). KRITON: A knowledge acquisition tool for expert systems. International Journal of Man-Machine Studies 26(1) 29-40.
**[Feldman et al., 1995]** Fldman R. and Hirsh H. Knowledge discovery in textual databases (KDT). In Proceedings of the First International Conference on Knowledge Discovery and Data Mining, Monreal, 1995.
**[Gomez et al., 1990]** Gomez, F. and Segami, C. (1990). Knowledge acquisition from natural language for expert systems based on classification
**[Hobbs, et al., 1992]** Hobbs, Jerry R., Douglas E. Appelt, John Bear, David Israel, Megumi Kameyama, and Mabry Tyson, 1992. ``FASTUS: A System for Extracting Information from Text'', Proceedings, Human Language Technology, Princeton, New Jersey, pp. 133-137, March 1993.

**[Murthy et al, 1994]** Murthy S.K., Kasif S., Salzberg S. A System for Induction of Oblique Decision Trees// Journal of Artificial Intelligence Research. N. 8, 1994.
**[Nahm et al., 2000]** Nahm U.Y., Mooney R.J. Using Information Extraction to Aid the Discovery of Prediction Rules from Text. Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining (KDD-2000) Workshop on Text Mining, pp. 51 - 58, Boston, MA, August, 2000.
**[Quinlan, 1986]** Quinlan J.R. Induction of Decision Trees // Machine Learning Journal N 1, 1986.
**[Quinlan, 1992]** Quinlan. C 4.5: Programs for Machine Learning. Morgan Kaufmann, Los Altos, CA, 1992. Problem-solving methods. Knowledge Acquisition 2(2) 107-128.
**[Tanabe et al., 1999]** Tanabe L., Scherf U., Smith L.H, Lee J.K., Hunter L., and Weinstein J.N. BioTechniques, V.27, N. 6, pp.1210-1217, December 1999.
**[TextAnalyst]** TextAnalyst tm, Microsystems, Ltd. http://www.analyst.ru/
**[Schreiber et al., 1988]** Schreiber G., Breuker, J., Bredweg B., Wielinga B. Modelling in KBS Development // Proceedings of the Second European Knowledge Acquisition Workshop (EKAW-88). Bonn, 1988. P.1-15.
**[Schreiber et al, 1993]** Schreiber, A.Th., Wielinga, B.J. and Breuker, J.A., Ed. (1993). KADS: A Principled Approach to Knowledge-based System Development. London, Academic Press.
**[Wolfgang, 2000]** Wolfgang May. An Integrated Architecture for Exploring, Wrapping, Mediating and Restructuring Information from the Web.Australasian Database Conference (ADC 2000). Jan. 31 - Feb. 3, 2000, Canberra, Australia. Australian Computer Science Communications, Vol. 2, No. 2, IEEE CS Press, p. 82-89. http://www.informatik.uni-freiburg.de/~dbis/Publications/2K/adc2k.html