УДК 624.012:801.52

# МЕТОДЫ ДЛЯ СЕМАНТИЧЕСКОГО ОТОБРАЖЕНИЯ ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ В МОДЕЛЯХ

М.Д.КУТУЕВ, А.Т.АБДЫШОВА
*E.mail. ksucta@elcat.kg*

*Бул маклада жаңы маалымат технологияларынын каражаттары менен техникалык тексттерди моделдештирүү жана оптимизациялоо проблемасы боюнча Европа өлкөлөрүндөгү (Франция, Бельфор, UTBM) жетишкендиктерди үйрөнүүнүн жана анализинин жыйынтыктары келтирилген.*

*В статье приводятся результаты изучения и анализа достижений европейских стран (Франция, Бельфор, UTBM) по проблеме моделирования и оптимизации технических текстов средствами новых информационных технологий.*

*This is article is devoted to the results of studying and analyze of achievements in European countries (France, Belfort, UTBM) by problems of modeling and optimization of technical texts with applied New Information Technologies.*

В машинной обработке текстовой информации роль памяти человека выполняет компьютерная система – онтология: именно она позволяет совместить анализ текста с его компьютерным «пониманием». Процедурно это достигается достаточно просто: необходимо найти проекцию текста на компьютерную онтологию. Текст рассматривается не только как вместилище информации – данных, фактов и знаний, которые требуется из него извлечь.

Решению задачи раскрытия семантического ресурса текста способствует Система семантического анализа Естественного Языка текстов, которая удовлетворяет следующим требованиям.

Первое. Партнеры интеллектуального общения вместе с текстом погружены в единую компьютерную среду онтологического знания.

Второе. Предварительная лингвистическая обработка исходного текста (морфологический, синтаксический и семантический анализ предложений) необходима для снятия «лексической оболочки» и выделения термов, несущих содержательную нагрузку.

Третье. Результатом компьютерного семантического анализа связного текста должен быть формальный или адаптированный текст Естественного Языка, который выражает его смысловое содержание.

Четвертое. Система должна обеспечивать самоконтроль авторского намерения – насколько адекватно он выражает свои мысли.

Пятое. Система должна многократно активизировать текст с целью более глубокого проникновения в смысл сообщения.

В результате самого общего взгляда на желаемые качества Системы семантического анализа можно сделать вывод, что потенциальные возможности текста реализуются при помощи двух механизмов: анализа через онтологию и активного диалога.

На рисунке показана блок-схема Системы семантического анализа Естественного Языка текстов, в которой взаимодействуют указанные выше функциональные компоненты.

Одним из ключевых понятий, характеризующим выбор метода анализа текстовой информации, а также реализацию конкретного варианта поиска, является модель поиска .

Модель поиска – это сочетание следующих составляющих: способа представления документов; способа представления поисковых запросов; вида критерия релевантности

документов. Вариации этих составляющих определяют большое число всевозможных реализаций систем текстового поиска.


**Блок-схема Системы семантического анализа ЕЯ текстов**

***Простейшие модели поиска***. Это модели, в которых документ представляется в виде набора ассоциированных с ним внешних атрибутов. К простейшим моделям поиска относятся модель дескрипторного поиска и модель, основанная на Дублинском ядре. В простейших системах дескрипторного поиска представление документа описывается совокупностью слов или словосочетаний лексики предметной области, которые характеризуют содержание документа. Эти слова и словосочетания называются дескрипторами.

*Дублинское ядро* (DublinCore) – это набор элементов метаданных, смысл которых зафиксирован в спецификации определяющего его стандарта. В терминах значений этих элементов можно описывать содержание различного рода текстовых документов. Первоначальная версия Дублинского ядра была предложена в 1995 году на состоявшемся в Дублине (США) симпозиуме, организованном OnlineComputerLibraryCenter (OCLC) и NationalCenterforSupercomutingApplications (NCSA) для описания информационных ресурсов библиотечных систем. В модели поиска, основанной на Дублинском ядре, представлением k-го документа является множество пар $D_k = \{(N_{ik}, V_{ik})\}$, где $N_{ik}$ – имя i-го элемента метаданных Дублинского ядра в описании содержания k-го документа; $V_{ik}$ – значение этого элемента метаданных. Представлением запроса также является множество пар некоторых элементов Дублинского ядра и их значений $Q = \{(N_j, V_j)\}$, где $N_j$ – имя j-го элемента метаданных Дублинского ядра в описании пользовательского запроса; $V_j$ – значение этого элемента метаданных. Критерий релевантности k-го документа выглядит следующим образом: $Q \subseteq D_k$.

***Модели, основанные на классификаторах***. Это одна из разновидностей простейших моделей поиска. Документ в данной модели представляется в виде совокупности ассоциированных с ним атрибутов. Атрибутами являются идентификаторы классов, к которым относится данный документ. Классы формируют иерархическую структуру классификатора. Запрос может быть представлен двумя способами:

- Простой вариант – запросом является идентификатор какого-либо класса из заданного классификатора. Критерий релевантности документа запросу – класс документа совпадает с классом в представлении запроса или является его подклассом.
- Сложный вариант – в запросе можно указать несколько классов классификатора. Критерий релевантности документа запросу – класс документа совпадает с каким-либо из указанных в запросе классов или является его подклассом.

***Векторные модели***. В настоящее время векторные модели являются самыми распространенными и применяемыми на практике моделями поиска. Векторные модели, в отличие от булевых, без труда позволяют ранжировать результирующее множество документов запроса. Суть таких моделей сводится к представлению документов и запросов в виде векторов. Каждому терму $t_i$ в документе $d_j$ и запросе q сопоставляется некоторый неотрицательный вес $w_{ij}$ ($w_i$ для запроса). Таким образом, каждый документ и запрос может быть представлен в виде k-мерного вектора: $\vec{d}_j \overset{def}{=} (w_{1j}, w_{2j}, \ldots, w_{kj})$, где k – общее количество различных термов во всех документах. Согласно векторной модели, близость документа $d_i$ к запросу q оценивается как корреляция между векторами их описаний. Вариации всевозможных способов назначения весов термов и оценки меры близости векторов определяют широкий спектр различных модификаций данной модели поиска.

***Вероятностные модели***. Впервые идеи таких моделей были предложены в 1960 году. В их основе лежит *принцип вероятностного ранжирования* (ProbabilisticRankingPrinciple – PRP). Этот принцип заключается в следующем: наивысшая общая эффективность поиска достигается в случае, когда результирующие документы ранжируются по убыванию вероятности их релевантности запросу. Существуют различные способы получения этих оценок, а также дополнительные предположения и гипотезы на основе априорных сведений относительно документов коллекции, которые и определяют конкретную реализацию вероятностной модели поиск.

***Сети вывода***. Так же, как и вероятностные модели, сети вывода основаны на принципе вероятностного ранжирования результирующих документов поиска. Главное их отличие от вероятностных моделей заключается в том, что используется оценка не вероятности релевантности документа запросу, а вероятности того, что он удовлетворяет информационным потребностям пользователя.

Сеть вывода основана на Байесовской сети, которая включает узлы четырех видов. Узлами первого вида являются документы коллекции, изученные пользователем в процессе поиска. Узлами второго вида являются термы, которыми описывается содержание документов. Узлами третьего вида являются запросы, состоящие из термов, которыми описывается содержание документов. Узел четвертого типа в сети только один, и он соответствует информационным потребностям пользователя, которые не известны поисковой системе. Все узлы первого и второго вида формируются заранее для заданной коллекции. Узлы третьего вида и их связи с узлами термов, описывающих документы, и узлом информационных потребностей формируются для каждого конкретного запроса. Методы тематического анализа текста можно разделить на две большие группы:

- лингвистический анализ;
- статистический анализ.

Первый ориентирован на извлечение смысла текста по его семантической структуре. Второй – по частотному распределению слов в тексте

***Лингвистический анализ***. Лингвистический анализ можно разделить на четыре взаимодополняющих анализа.

– ***Лексический анализ***. Заключатся в разборе текстовой информации на отдельные абзацы, предложения, слова, определении национального языка изложения, типа предложения, выявлении типа лексических выражений (бранных, жаргонных слов) и т.д. Данный вид анализа не представляет существенной сложности для реализации.

– ***Морфологический анализ***. Сводится к автоматическому распознаванию частей речи каждого слова текста (каждому слову ставится в соответствие лексико-грамматический класс). Часто морфологический анализ используется в статистических методах анализа при предварительной процедуре обработки документов – приведении слов к базовой форме.

– ***Синтаксический анализ***. Заключатся в автоматическом выделении

семантических элементов предложения: именных групп, терминологических целых, предикативных основ. Это позволяет повысить интеллектуальность процесса обработки тестовой информации на основе обеспечения работы с более обобщенными семантическими элементами.

– ***Семантический анализ***. Заключается в определении информативности текстовой информации и выделении информационно-логической основы текста. Проведение автоматизированного семантического анализа текста предполагает решение задачи выявления и оценки смыслового содержания текста.

***Статистический анализ.*** Статистический анализ – это, как правило, частотный анализ в тех или иных его вариациях. Общая суть такого анализа заключается в подсчете количества повторений слов в тексте и использовании результатов подсчета для конкретных целей. Рассмотрим один из наиболее эффективных статистических подходов.

***Латентно-семантический анализ***. Латентно-семантический анализ (LSA – LatentSemanticAnalysis) – это теория и метод для извлечения контекстно-зависимых значений слов при помощи статистической обработки больших наборов текстовых данных. Данный метод анализа используется не только в области поиска информации, но и в задачах фильтрации и классификации. Один из самых распространенных вариантов LSA основан на использовании разложения исходной матрицы по сингулярным значениям (SVD – Singular-ValueDecomposition). Согласно теореме о сингулярном разложении, любая вещественная прямоугольная матрица X может быть разложена в произведение трех матриц: $X = U\Sigma V^T$, где матрицы U и V – ортогональные, а $\Sigma$ – диагональная матрица, значения на диагонали которой называются сингулярными значениями матрицы X. Идея такого разложения и суть латентно-семантического анализа заключаются в том, что если в качестве X использовалась матрица термов на документ, то матрица $\hat{X}$, содержащая только k первых линейно независимых компонент X, отражает основную структуру ассоциативных зависимостей, присутствующих в исходной матрице, и в то же время не содержит шума. Таким образом, каждый терм и документ представляется при помощи векторов в общем пространстве размерности k (так называемом *пространстве гипотез*).

Из предложенных моделей различных версий и интерпретаций наиболее интересным и перспективным с точки зрения использования его в технических текстах, по нашему мнению, является вероятностный подход. Так как технические тексты в реальности являются недетерминированными, остальные модели тоже имеют определенные положительные моменты и применяются, в основном, в гуманитарных науках.

## Список литературы


1. Адарюков В.И. Исследование и разработка машинно-ориентированного метода инфологического моделирования информационно-поисковых систем фактографического типа: Диссертационная работа к.т.н.: 05.13.06 / Ленинградский электротехнический институт им В.И. Ульянова (Ленина). – СПб., 1988. – 256 с.
2. Сэлтон Г. Автоматическая обработка, хранение и поиск информации: Пер. с англ. / Под ред. А.И. Китова. – М.: Советское радио, 1973. – 560 с.
3. Белянин В.П. Введение в психолингвистику. – Изд. 2-е, испр. и доп., – М.: ЧеРо, 2000. – 128 с.
4. Боровиков В. STATISTICA. Искусство анализа данных на компьютере: Для профессионалов. 2-е изд. – СПб.: Питер, 2003. – 688 с.
5. Добрынин В.Ю., Некрестьянов И.С., Задача выбора тематических коллекций, релевантных запросу. // Труды Всероссийской научно-методической конференции "Интернет и современное сообщество". –, Санкт-Петербург., декабрь 1998.

6. Некрестьянов И.С. Тематико-ориентированные методы информационного поиска: Диссертационная работа к.т.н.: 05.13.11 / Санкт-Петербургский государственный университет – СПб., 2000. – 80 с.

7. Harman D. Latent semantic indexing (LSI) and TREC-2. In Proc. of the Second Text REtrieval Conference, 1994.

8. Дубинский А.Г. Некоторые вопросы применения векторной модели представления документов в информационном поиске // Управляющие системы и машины. - – 2001. -– № 4. - – С. 77-83.

9. Ермаков А.Е. Полнотекстовый поиск: проблемы и их решение // Мир ПК. – 2000. –- N№ 5.

10. Ермаков А.Е. Неполный синтаксический анализ текста в информационно-поисковых системах // Компьютерная лингвистика и интеллектуальные технологии: Труды Международного семинара «Диалог’2002». В двух томах. Т.2. "Прикладные проблемы". – Москва., 2002. - – С. 180-185.

11. Ермаков А.Е., Плешко В.В. Ассоциативная модель порождения текста в задаче классификации // Информационные технологии. –- 2000. –- N№ 12.

12. Иванов В., Некрестьянов И., Пантелеева Н. Расширение представления документов при поиске в Веб // Труды четвертой всероссийской конференция RCDL'2002. В двух томах. Т.2. –- Дубна, 2002. –- С. 55-68.

13. Когаловский М. Р. Перспективные технологии информационных систем. – М.: ДМК Пресс; М.: Компания АйТи, 2003. – 288 с.

14. Когаловский М.Р. Энциклопедия технологий бах данных. – М.: Финансы и статистика, 2002. – 800 с.

15. Кураленок И.Е., Некрестьянов И.С. Оценка систем текстового поиска // Программирование. – 2002. – N№ 4. – С. 226-242.

16. Кураленок И.Е., Некрестьянов И.С. Автоматическая классификация документов на основе латентно-семантического анализа // Труды первой всероссийской научно-методической конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции". – СПб., 1999. - – С. 89-96.