Роль искусственных языков при контекстно-свободном поиске в массиве текстов и поиске с использованием метаданных

The Role of Artificial Languages in Context-free Search of Text Massifs and Search Using Metadata

Роль штучних мов при контекстно-вільному пошуку в масиві текстів і пошуку з використанням метаданих

Соколовский В. В.
Государственная публичная научно-техническая библиотека России, Москва, Россия

Vladimir Sokolovsky
Russian National Public Library for Science and Technology, Moscow, Russia

Соколовський В. В.
Державна публічна науково-технічна бібліотека Росії, Москва, Росія

Приводятся аналогии и различия в развитии методов контекстно-свободного поиска и поиска с использованием метаданных. Рассматриваются подходы, позволяющие формализовать текстовые документы на естественном языке, чтобы представить их в компьютерно-обрабатываемом виде для целей поиска.

The analogies and differences in developing the methods of context-free searching and searching using metadata are discussed. The approaches towards formalization of text documents in natural language for their representation in machine-readable formats for searching purposes are considered.

Наводяться аналогії та відмінності в розвитку методів контекстно-вільного пошуку та пошуку з використанням метаданих. Розглядаються підходи, що дозволяють формалізувати текстові документи природною мовою, щоб представити їх в придатному для комп'ютерної обробки вигляді для цілей пошуку.

Контекстно-свободный поиск информации является одной из наиболее актуальных задач, решаемых с помощью методов семантического анализа. Постановку этой задачи можно формализовать как нахождение всех текстов из некоего массива, написанных на естественном языке и «похожих» на заданный текст-образец. В качестве характеристики подобия текстов между собой, выбирается способ (включая и выбор формулы) для подсчёта численной меры подобия. Обычно текстовые документы считаются в той мере подобными друг другу, в какой подобен их терминологический состав.

Текстовые документы на естественном языке можно формализовать в другом виде. С 1968 года [1, 233 с.] начал развиваться подход, предлагающий использовать для представления смысла предложений на естественном языке падежные фреймы (case frame). При разборе предложения, глагол выбирается за основу фрейма, а глагол связывается с *агентом* (тем, кто совершает действие), *объектом*, *местом*, *временем* действия и т.п. Таким образом, предложение, представленное падежным фреймом, также представляет собой граф, в узлах которого находятся объекты, а рёбра графа представляют собой отношения (из некоторого словаря отношений) между этими объектами.

Например, подход, связанный с автоматическим преобразованием текста на естественном языке к формальному графовому виду, практикует рабочая группа Aot.ru [2]. Согласно этому подходу, грамотная декомпиляция языковых механизмов позволит максимально приблизить человеческий язык к современному компьютеру. То есть, цель – сделать информацию машинно-обрабатываемой. Рабочая группа Aot.ru разрабатывает программное обеспечение в области автоматической обработки текста, в основном связанное с анализом русского языка. Один из их проектов, семантический анализ текста на русском языке, представляет собой построение семантического графа текста. Семантический анализ строит семантическую структуру одного предложения на русском языке. Семантическая структура состоит из семантических узлов и семантических отношений.

Мера подобия текстов, основанная на частотном анализе терминов, входящих в текстовый документ хороша своей универсальностью, в отличие от меры подобия, основанной на анализе падежных фреймов или графов текста, использование которой накладывает ограничения на размеры и прочие характеристики текстов [3]. Также, частотный анализ терминов обладает меньшей требовательностью к вычислительным ресурсам компьютера, чем синтаксический и семантический разбор предложений. Зато, в тех случаях, когда анализ падежных фреймов текста применим, этот подход позволяет увеличить точность поиска.

Другим направлением развития методов поиска, в отличие от контекстно-свободного поиска, является поиск по метаданным, описывающим смысл текста (например, ключевые слова, рубрики, аннотация). Ключевые слова, рубрики, аннотация и другие метаданные, описывающие документ, составляются экспертами в описываемой области, или даже автором документа. Качество метаданных, созданных экспертами в описываемой области пока является недостижимым с помощью программ семантического анализа текста. В настоящее время, в общем случае, метаданные, создаваемые экспертами, описывают ресурс более качественно и точно, чем результаты применения процедур автоматического создания метаданных с использованием семантического анализа полнотекстовых источников. Поэтому достижимая при этом точность поиска по таким метаданным должно быть выше, чем точность поиска по полнотекстовым источникам.

В отличие от ключевых слов и рубрик, аннотация, обычно являющаяся текстом на естественном языке, сложнее подвергается компьютерной обработке. Аннотация является метаданными, и в этом смысле представляет собой инструмент для осуществления качественного поиска. Но поскольку аннотация представляет собой текст на естественном языке, то для компьютерной обработки к ней необходимо применять методы обработки текстов на естественном языке. А именно, либо работать с частотами терминов, встречающихся в аннотации, либо разбирать аннотацию на падежные фреймы, со всеми вытекающими плюсами и минусами каждого подхода, которые были рассмотрены в начале статьи. Существует третий вариант, позволяющий избежать основных недостатков первых двух подходов. Таким недостатком является неточность, вытекающая из статистической природы методов, опирающихся на анализ частот терминов, содержащихся в тексте. А при разборе аннотации на падежные фреймы, недостатком являются ошибки, связанные с несовершенством методов «понимания» текстов на естественном языке. Этот третий вариант работы с аннотациями заключается в том, чтобы аннотация не преобразовывалась с естественного языка к машинно-обрабатываемой форме, а сразу составлялась бы на искусственном, строго формализованном языке. Подходящие для этого языки уже разработаны в достаточной мере.

Работы по представлению структур понятий и ассоциаций в виде графов и семантических сетей велись с начала XX века [1, 230 с.]. Работы в этой области показали силу графов для моделирования ассоциативного смысла, но были ограничены чрезмерной общностью формализма – были формализованы только самые общие отношения. Исследования в области сетевых представлений часто фокусировались на спецификации этих отношений. Само по себе представление отношений в виде графов имеет мало преимуществ перед исчислением предикатов – это только другая запись отношений между объектами. Сила сетевых представлений состоит в определении связей и специфических правил вывода, определяемых механизмом наследования. За счёт реализации базовых семантических отношений как части формализма, а не как части знаний о предметной области, базы знаний позволяют автоматизировать работу и обеспечить большую общность и непротиворечивость. Структуры, в которых базовые семантические отношения включили в себя семантические отношения слов в естественном языке, были названы, как уже упоминалось, падежными фреймами. Фреймы – это схема представления, ориентированная на включение в строго организованные структуры данных неявных (подразумеваемых) информационных связей, существующих в предметной области. Фреймы расширяют возможности семантических сетей, позволяя представлять сложные объекты не в виде семантической структуры, а в виде единой сущности (фрейма). Это также позволяет естественным образом представить стереотипные сущности, классы, наследование и значения по умолчанию. Исследования этих идей привели к разработке философии объектно-ориентированного программирования. С 80-х годов XX века начались разработки сетевых языков для моделирования различных предметных областей. Один из таких языков называется концептуальными графами. В настоящее время развитие языков, позволяющих описывать предметную

область активно продолжается и в отношении мощности языка, правил вывода, а также в отношении развития программных систем для практического применения таких языков.

Литература

1. Люггер, Джордж, Ф. Искусственный интеллект: стратегии и методы решения сложных проблем, 4-е издание. : Пер. с англ. – М. : Издательский дом «Вильямс», 2005. – 864 с.
2. Автоматическая Обработка Текста. // http://www.aot.ru
3. Соколовский В. В. Исследование качества автоматической классификации текстовых документов с использованием семантического графа документа. // Автоматизированные библиотечные системы и технологии: Материалы X Международной конференции «Крым 2005».
4. Рассел Стюарт, Норвинг Питер, Искусственный интеллект: современный подход, 2-е изд.: Пер. с англ. – М.: Издательский дом «Вильямс», 2006. 1408 с.
5. Базы знаний и интеллектуальных систем / Т.А.Гаврилова, В.Ф.Хорошевский.– СПб.: Питер, 2001.–384 с.