Манучарян Л.А.
аспирант третьего года обучения
Воронежская государственная лесотехническая академия
Российская Федерация, г.Воронеж

ПРИМЕНЕНИЕ СИСТЕМ ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ В НАУКОЕМКИХ ИНДУСТРИЯХ

Системы извлечения информации могут полезны для решения ряда задач в наукоемкой индустрии. Одним из примеров эффективности таких систем является предоставление возможности проведения мониторинга информационных ресурсов, доступных в глобальной сети, что является важным фактором конкурентоспособности предприятий любого типа. Система извлечения должна включать инструментарий для идентификации концепций и понятий, соответствующих интересам пользователя, и предоставить возможность сделать мониторинг по выбранным направлениям. Более продвинутый вариант системы извлечения должен обеспечить следующие возможности:

1) управление и модификация онтологий [1], на основании результатов извлечения информации из глобальной сети для заданных целевых знаний.
2) отслеживание изменении тенденции в интересующих направлениях, с предупреждением об изменениях.
3) Модель создания общих и целевых поисковых агентов, которые могут использовать онтологии, для поиска информации из разных онлайн источников.
4) Платформа для интеграции информации из разных источников, а также объединение, анализ и публикация этой информации.

В будущем, системы извлечения могут быть расширены для поддержки большего числа типов индустрий.

Системы извлечения являются основой создания систем управления базами знаний (далее, СУБЗ). На данный момент, распространение СУБЗ стало переломным моментом в наукоемкой индустрии. Такие системы

внедрены в стратегические, политические и исполнительные процессы в институтах и организациях по всему миру. Глобальный рынок СУБЗ вырос вдвое, по сравнению с 1991 годом и в 2010-ом оборот превысил 10 миллиард долларов США. Ожидается, что применение СУБЗ сэкономит Топ 500 компаниям мира примерно 31 милиард долларов США в 2012-ом.

Направление занятости (трудоустройства) является одним из наиболее исследуемых областей в плане управления базами знаний, так как каждое предприятие должно учесть этот момент. Отделы кадров в предприятиях всегда нуждаются в СУБЗ, для мониторинга соответствующих изменений в отрасли, а многие сторонние компании-консультанты по трудоустройству используют эту возможность для получения сведений о любых изменениях в рынке человеческих ресурсов. Существует много онлайн систем поиска работы, в которых интенсивно используется СУБЗ, для эффективного вычисления критериев соответствия работников с требованиями работодателей.

Область трудоустройства может быть эффективного использована в системах СУБЗ, так как содержит много общих типов концепций, что означает легкость адаптирования общей системы под данную область, и, второе, для сопровождения системы не требуются специалисты по данной области знаний, для осмысления использованных понятий и концепции. Таким образом, вся работа может быть выполнена разработчиком без особых знаний по конкретной области. Эти два момента очень важны для быстрой разработки СУБЗ [2].

Рекламные объявления о вакансиях являются наилучшим индикатором изменений в индустрии. Посредством мониторинга этих объявлений за какой-то период времени, можно сделать заключения, например, об изменениях в требованиях каких-то навыков и в типах требуемой квалификации, о колебаниях средней зарплаты, распределении спроса на конкретную квалификацию, и так далее...

Извлечение информации, основанное на онтологиях.

Растущее число инструментов и ресурсов для Семантической Сети создает новые проблемы в области извлечения смысловой информации (далее, ИСИ), и, в частности, извлечения информации, базированного на онтологиях (далее, ИСИО). Одним из важных отличий традиционного ИСИ от ИСИО является использование формальных онтологий, вместо плоского словаря, а также возможность применения логических умозаключений. Другим различием является факт, что ИСИО не только находит (самый специфический) тип извлеченной сущности, но также идентифицирует его, ассоциируя с семантическим определением в онтологии. Это позволяет обнаружить сущности параллельно в множестве документов, и сделать обогащение описаний сущностей, в процессе извлечения. Если онтология уже наполнена соответствующими экземплярами, задачей ИСИО является простое распознавание экземпляров онтологии в тексте. В отличие от

традиционных ИСИ, для которых наборы обучения существуют в огромных количествах, для семантических веб-приложений чувствуется недостаток существующих материалов на данный момент. Новые наборы обучения должны создаваться вручную или полуавтоматическим образом, что является довольно трудоемкой работой, несмотря на то, что на данный момент разрабатываются системы по созданию таких наборов [3]. Одним из преимуществ ИСИО по сравнению с ИСИ, является факт, что в первом, выходные данные (семантические метаданные о тексте) связываются с онтологией, что позволяет извлечь намного больше полезной информации из текста, посредством, например, использования "родственной" информации, или применения логических заключений. Это позволяет получить более полное представление о тексте и сделать более полезные заключения. Например, в случае с областью трудоустройства, определения мест, где есть вакансии, задача легкая для обеих типов систем (ИСИ и ИСИО), однако, связывание также городов и стран (ИСИО) позволяет иметь намного больше полезной информации, из-за обеспечения возможности проведения анализа для конкретных регионов, например, что индустрия информационных технологий процветает в Новосибирске, или же что в Москве и Московской области предлагаются лучшие варианты соц. пакетов, чем в других областях Российской Федерации.

Использованные источники:

1. Онтологии. URL: http://ru.wikipedia.org/wiki/%D0%9E%D0%BD%D1%82%D0%BE%D0%BB%D0%BE%D0%B3%D0%B8%D1%8F_(%D0%B8%D0%BD%D1%84%D0%BE%D1%80%D0%BC%D0%B0%D1%82%D0%B8%D0%BA%D0%B0)

2. D. Maynard, V. Tablan, H. Cunningham, C. Ursu, H. Saggion, K. Bontcheva, and Y. Wilks. Architectural Elements of Language Engineering Robustness. Journal of Natural Language Engineering – Special Issue on Robust Methods in Analysis of Natural Language Data, 8(2/3):257–274, 2002.

3. B. Popov, A. Kiryakov, A. Kirilov, D. Manov, D. Ognyanoff, and M. Goranov. KIM Semantic Annotation Platform. In 2nd International Semantic Web Conference (ISWC2003), pages 484–499, Berlin, 2003. Springer.