УДК 004.89

ВЫЯВЛЕНИЕ НОВЫХ ТЕХНОЛОГИЧЕСКИХ ТРЕНДОВ: ПРОБЛЕМЫ И ПЕРСПЕКТИВЫ

В.Ф. Хорошевский (*khor@ccas.ru*)
ВЦ РАН, НИУ ВШЭ, Москва

В докладе обсуждаются вопросы автоматизации процессов выявления новых технологических трендов на основе обработки документов разных жанров. Представлен гибридный подход к выявлению новых технологических трендов, где для повышения качества результатов используются методы и средства статистической обработки коллекций документов, интегрированные с методами извлечения информации из текстов.

Введение

Общепризнанной «горячей» точкой в исследованиях и разработках по тенденциям научно-технического прогресса в настоящее время является выявление новых технологических трендов. На современном этапе работы в данной области концентрируются, в основном, на построении дорожных технологических карт с использованием методологий Форсайт-прогнозирования и методик Дельфи, а основные активности сосредоточены в части опросов экспертов, последующей статистической обработке полученных результатов и на подготовке людьми-экспертами отчетов, в которых представлены построенные дорожные карты с соответствующими пояснениями [Переслегин, 2009; Bagheri et al., 2009]. При этом наиболее проработанными (в смысле автоматизации) являются этапы формирования базы данных с результатами анкетирования и методы их математической обработки.

Не отрицая важности и сложности автоматизации указанных этапов, в настоящей работе акцент делается на исследовании процессов выявления новых технологических трендов на основе обработки электронных документов разных жанров с целью последующей генерации опросников для экспертов. Методы и средства автоматизированного формирования экспертных групп и выявления центров экспертной компетенции в определенных предметных областях представлены в работе [Хорошевский, 2010].

Изложение материала структурировано следующим образом. В первом разделе обозначен спектр литературы, где обсуждаются различные подходы к решению задачи выявления новых технологических трендов и показывается, что перспективные решения в данном случае целесообразно рассматривать в контексте использования семантических технологий. В следующем разделе предлагается гибридный подход к решению вышеуказанной задачи, в рамках которого основное внимание уделяется онтологическому моделированию процессов выявления новых технологических трендов, что, в свою очередь, является базисом для использования методов и средств статистической обработки коллекций документов, интегрированных с методами извлечения информации из текстов. В заключительной части работы обсуждаются направления дальнейших исследований и разработок.

1. Выявление новых технологических трендов: состояние работ и основные проблемы

Анализ работ в области выявления технологических трендов [Daim et al., 2006; Shibata et al., 2008; Bagheri et al., 2009; Youngho et al., 2009; Wang et al., 2010] показывает, что новые подходы в данной области все больше связываются с интеграцией классических методов прогнозирования на базе соответствующих методик опроса экспертов и гибридных методов автоматической обработки корпусов текстов с использованием статистических методов и методов извлечения информации из текстов.

При этом основные проблемы автоматизации выявления новых технологических трендов связаны с тем, что в рамках обработки соответствующих информационных ресурсов необходимо

- различать тексты разных жанров;
- для каждого из жанров использовать собственные модели извлечения информации;
- интегрировать полученные частные результаты в рамках единой модели представления знаний.

С учетом вышесказанного представляется, что наиболее перспективным в данном случае является онтологическое моделирование процессов выявления новых технологических трендов и автоматизированная обработка информационных ресурсов на основе гибридного подхода, в рамках которого естественным образом объединяются статистические и лингвистические методы.

По существу, при этом лингвистические методы обеспечивают автоматическое формирование характеристических векторов обрабатываемых текстов, а статистические методы используются для автоматической кластеризации коллекций документов на основе TF-IDF подхода.

2. Гибридный подход к выявлению новых технологических трендов

2.1. Предварительные замечания

В настоящей работе для выявления новых технологических трендов используется гибридный подход, который, с одной стороны, развивает методы, недавно предложенные в работах [Youngho et al., 2009; Wang et al., 2010], а с другой – обобщает их на случай мультиязычных коллекций документов различных жанров с активным использованием онтологических моделей, под управлением которых осуществляется автоматическое извлечение информации из тестов и формирование не просто “bag of words” для каждого текста, но обогащение характеристических векторов важными для предметной области ключевыми выражениями, которые формируются за счет использования специальных паттернов. Спецификой обсуждаемого дальше подхода является и то, что после предварительной статистической обработки текстов происходит автоматическая генерация OWL-представления экземплярной части онтологической модели тренда, предложенной И.В. Ефименко, а также объединение (merging) OWL-представлений для коллекций документов одного жанра и объединение OWL-представлений для коллекций разных жанров. Ниже указанные аспекты обсуждаются подробнее.

2.2. Базовые гипотезы

В настоящей работе для выявления новых технологических трендов предлагаются следующие базовые гипотезы:

1. В качестве модели прогнозирования использование кривых Гартнера [GARTNER, 2012], где явно выделяются области «Технологический триггер», «Пик завышенных ожиданий», «Ущелье утраты иллюзий», «Склон осознания» и «Плато продуктивности».
2. Для анализа информации на уровне «Технологического триггера» использование коллекций научно-технических публикаций в области охвата прогнозируемого тренда.
3. Для уровней «Пика завышенных ожиданий» и «Ущелья утраты иллюзий» использование новостных сайтов по тематике исследуемого тренда, обработка которых, как правило, фиксирует всплеск интереса к новым технологическим трендам.
4. Для обработки информации на уровнях «Склона осознания» и «Плато продуктивности» предлагается осуществлять патентный анализ в области охвата прогнозируемого тренда.
5. Интеграцию результатов обработки коллекций отдельных жанров предлагается проводить на основе пересечения и/или объединения результатов статистической обработки отдельных коллекций.

2.3. Используемые онтологические модели

Как отмечалось выше, в настоящей работе используется онтологическая модель технологических трендов, разрабатываемая в настоящее время в рамках государственного контракта № 07.524.12.4018 «Исследование и разработка моделей долгосрочного технологического прогнозирования и программного комплекса Интерактивная дорожная карта с обратной связью». В силу ограничений на объем настоящей работы онтология технологических трендов здесь не рассматривается, хотя другие онтологические модели, которые кратко рассматриваются ниже, на нее опираются.

В дополнение к онтологии технологических трендов в рамках реализации гибридного подхода к их выявлению используется онтологическая модель OWL-представления трендов (Рис. 1), специфицированная в системе Protégé [Protégé, 2012].

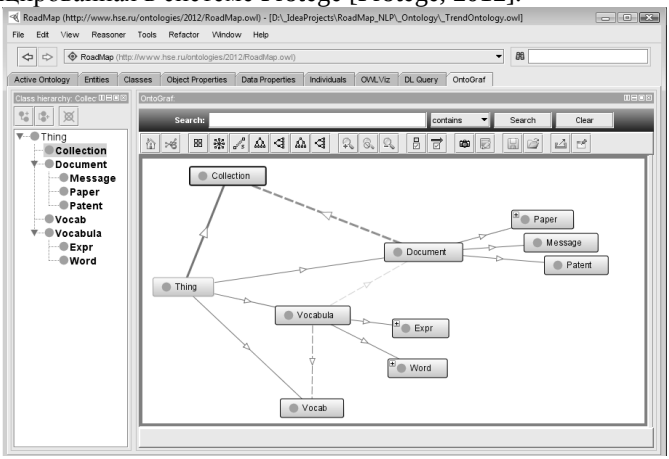


Рис.1. Онтология OWL-представления трендов в системе Protege

Приведенная онтология используется OWL-генератором экземплярной части онтологической модели тренда.

2.4. Реализация гибридного подхода к выявлению новых технологических трендов

Как показывает наш опыт создания систем извлечения информации из многоязычных коллекций документов [Efimenko, et al. 2009; Хорошевский, 2011], а также указанные выше цели разработки гибридного модуля формирования характеристических векторов текстов различных жанров, в данном случае целесообразно использовать следующую совокупность программных ресурсов обработки текстов: лексическое форматирование + морфологизация + словарное означивание

+ извлечение простых именных групп + формирование характеристических векоров + генерация OWL-представления.

В качестве инструментария для извлечения информации из текстов в настоящей работе используется платформа GATE, расширенная плагинами NP Chunker и Russian Morph Tagger, в котором используется открытая версия модуля русской морфологии компании Яндекс, а также специально разработанным ресурсом формирования TF-IDF и ресурсом OWL Generator, который базируется на идеях, представленных в работе [Witte, et al., 2010].

3. Предварительные результаты

3.1. Коллекции документов

Для проверки базовых гипотез, перечисленных в подразделе 2.2 настоящей работы, были сформированы три коллекции документов из предметной области «Green Energy». Первая коллекция включала случайным образом выбранные из трудов международных конференций научные статьи, вторая коллекция – набор новостей по тематике данной предметной области из блогов и с сайтов коммерческих компаний, а третья коллекция содержала аннотации соответствующих патентов, взятых с сайта EPO (European Patent Office).

Каждая из коллекций документов была планаризована для дальнейшей обработки, каждый документ каждой коллекции был обработан с помощью реализованного нами гибридного модуля формирования характеристических векторов, а результаты были обработаны реализованным нами модулем формирования TF-IDF для коллекций документов. Полученные результаты поступали на вход реализованного нами модуля OWL Merger, на выходе которого формировалась экземплярная часть OWL-представления коллекции.

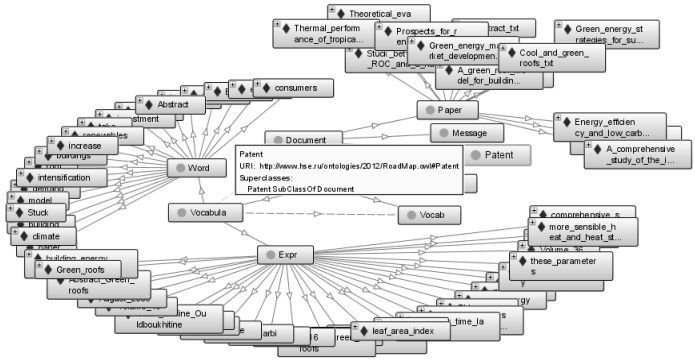
Валидация полученных результатов осуществляется в системе Protégé, которая в данном случае используется в качестве инсрументария онтологического инжиниринга.

3.2. Результаты выявления трендов в предметной области «Green Energy»

После обработки всех коллекций в соответствии со схемой, представленной выше, были получены результаты, показанные на Рис. 2, где для примера приведено OWL-представление коллекции научных статей.

OWL-представление новостной коллекции и результаты обработки патентов, к сожалению, плохо видны в мелком формате и потому исключены из текста настоящей работы.

Рис.2. OWL-представление коллекции научных статей в системе Protege



3.3. Анализ полученных результатов

Как показывает анализ первых полученных результатов, наименее интересные для дальнейшего использования в рамках автоматической генерации вопросников для экспертов получаются при обработке коллекции научных статей. На наш взгляд, данная ситуация определяется тем, что терминология предметной области активно расширяется с одновременным формированием своих систем терминов в разных научных школах. Поэтому представляется важным провести мониторинг более широкого спектра научных публикаций в данной области.

Интересными представляются результаты, полученные путем объединения коллекций документов разных жанров (Рис. 3). При этом в результирующее OWL-представление вошли только те термины, которые активно используются во всех коллекциях.

Следует отметить, что OWL-представление объединения коллекций документов разных жанров дает более полезные для дальнейшего использования результаты. Это объяняется, на наш взгляд, тем, что в таком случае наиболее релевантные термины и терминологические группы получают большие TF-IDF веса. Вместе с тем, и в данном случае нельзя утверждать, что в процессе автоматической обработки получены результаты, которые находятся на уровне и/или превышают результаты работы квалифицированных экспертов.

Поэтому направления дальнейших исследований и разработок в области автоматизации выявления новых технологических трендов должны быть связаны, прежде всего, с увеличением объема текстовых коллекций и их предварительного анализа с помощью методов корпусной лингвистики для накопления новых лингвистических шаблонов, формированием для дальнейшего использования представительных стоп-

словарей, а также подбором порогов TF-IDF для коллекций разных жанров.

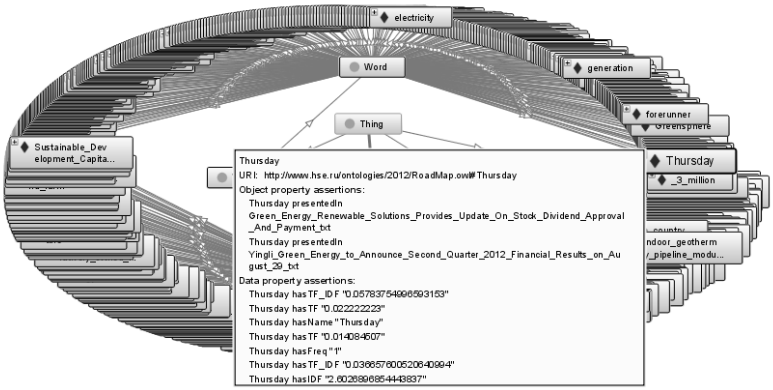


Рис.3. OWL-представление результатов объединения коллекций документов

Как представляется, такие модификации могут обеспечить приемлемое качество автоматического выявления новых технологических трендов.

Заключение

В настоящей работе представлены различные подходы к решению задачи выявления новых технологических трендов и предложен гибридный подход к решению вышеуказанной задачи, в рамках которого основное внимание уделено использованию методов и средств статистической обработки коллекций документов, интегрированных с методами извлечения информации из текстов. Проведен анализ первых результатов автоматического выявления технологических трендов и намечены направления дальнейших исследований и разработок.

Благодарности. Работа выполняется НИУ ВШЭ по заказу Министерства образования и науки РФ в рамках государственного контракта № 07.524.12.4018 «Исследование и разработка моделей долгосрочного технологического прогнозирования и программного комплекса Интерактивная дорожная карта с обратной связью».

Список литературы

[Переслегин, 2009] Переслегин С., Новые карты будущего или Анти-Рэнд. СПб.: Terra Fantastica, 2009.

[Хорошевский, 2010] Хорошевский В.Ф., Извлечение информации из текстов на конференциях серии ДИАЛОГ: взгляд соседа по лестничной клетке. // Труды международной конференции "Диалог 2010" М. Наука – 2010.

[Хорошевский, 2011] Хорошевский В.Ф., Пространства знаний в сети Интернет и Semantic Web (Часть 3), Искусственный Интеллект и Принятие решений, № 2 (2011).

[Bagheri et al., 2009] Bagheri S. K., Nilforoushan H., Rezapour M., Rashtchi M., A new approach to Technology Roadmapping in the Open Innovation context: The Case of Membrane Technology for RIPI, Journal of Science & Technology Policy, Vol. 2, N 1, Spring 2009.

[Daim et al., 2006] T.U. Daim, G. Rueda, H. Martin, Forecasting emerging technologies: use of bibliometrics and patent analysis, Technol. Forecast. Soc. Change, vol. 73, N 8, 2006.

[**Efimenko, et al. 2009**] Efimenko I., Minor S., Starostin A., Drobyazko G., Khoroshevsky V., Generating Semantic Content for the Next Generation Web, Chapter in Monograph "Semantic Web", Publisher IN-TECH, 2009, ISBN 978-953-7619-33-6

[**GARTNER, 2012**] Gartner home page, URL: http://www.gartner.com/technology/research.jsp

[Protégé, 2012] Protege Homepage, http://protege.stanford.edu/

[Shibata et al., 2008] Shibata et al. Detecting emerging research fronts based on topological measures in citation networks of scientific publications, Technovation, N 28 (2008).

[Wang et al., 2010] Wang et al. Identifying technology trends for RD planning using TRIZ and text mining, RD Management, vol. 40, N 5, 2010.

[Witte, et al. 2010] Witte R., Khamis N., Rilling J., Flexible ontology population from text: The owlexporter. In International Conference on Language Resources and Evaluation (LREC), Valletta, Malta, 05/2010 2010.

[Youngho et al., 2009] Youngho Kim, Yingshi Tian, Yoonjae Jeong, Ryu Jihee, Sung-Hyon Myaeng, Automatic Discovery of Technology Trends from Patent Text, In: Proc. SAC'09, March 8-12, 2009, Honolulu, Hawaii, U.S.A., 2009.