OSTIS-2013
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822:514

АВТОМАТИЗАЦИЯ ПРОЦЕССОВ ВЫЯВЛЕНИЯ ТЕХНОЛОГИЧЕСКИХ ТРЕНДОВ В СИСТЕМЕ АРМ «ТРЕНД»

Хорошевский В.Ф.

Федеральное государственное бюджетное учреждение науки, Вычислительный центр им. А.А. Дородницына РАН,
Центр информационно-аналитических систем ИСИЭЗ НИУ ВШЭ
г. Москва, Россия

khor@ccas.ru

vkhoroshevsky@hse.ru

В работе обсуждается АРМ «Тренд» - автоматизированное рабочее место поддержки процессов выявления новых технологических трендов на основе гибридного подхода к извлечению информации из текстов публикаций. Приводятся результаты обработки представительной коллекции документов, в которых представлены аннотации научно-технических статей.
Ключевые слова: технологический тренд; гибридный подход; лингво-статистический метод извлечения информации из текстов; автоматизированное рабочее место.

Введение

Общепризнанной «горячей» точкой в анализе тенденций 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, а также построению паттернов данных и анализу временных рядов, специфицирующих существующие и прогнозируемые тенденции.

Настоящая работа концентрируется на обсуждении вопросов автоматизации процессов выявления новых технологических трендов на основе гибридного подхода к обработке документов, ориентированного на интеграцию классических методов прогнозирования и гибридных методов автоматической обработки корпусов текстов с использованием статистических методов и методов извлечения информации из текстов [Glance, et al., 2004; Daim et al., 2006; Shibata et al., 2008; Bagheri et al., 2009; Kim, et al., 2009; Wang et al., 2010].

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 и соответствующей системы алгоритмов постобработки.

С учетом вышесказанного в настоящем исследовании для выявления новых технологических трендов предлагается использовать гибридный подход, который, с одной стороны, развивает методы, предложенные в работах [Kim, 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 предлагаемого подхода является и то, что после предварительной статистической обработки текстов происходит автоматическое объединение полученных представлений для коллекций документов одного жанра, а также генерация OWL-представления экземплярной части онтологической модели тренда.

1. Реализация гибридного подхода к выявлению трендов

1.1. Базовые гипотезы

В настоящем исследовании в процессе выявления новых технологических трендов используются следующие базовые гипотезы:

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

1.2. Онтологические модели технологических трендов

Как известно, **тренд** (от англ. Trend) это долговременная общая тенденция изменения исследуемого временного ряда [Wiki, 2012]. Таким образом, базовым понятием тренда является понятие **тенденции** (от ср.-век. лат. tendentia - направленность), как направления развития какого-либо явления, мысли, идеи... При этом в модель тренда естественным образом вовлекается понятие **процесса** (действия по знач. глагола «направлять»), что, в свою очередь, предполагает наличие совокупности последовательных действий, направленных на достижение определенного результата. При этом, с учетом целей и задач настоящей работы, понятие тренда целесообразно конкретизировать следующим образом: **технологический тренд** – активно развивающееся в последние 5 лет технологическое направление, которое, как ожидается, продолжит свое активное развитие в ближайшие 10 лет.

С учетом этого в модель технологического тренда вводятся следующие дополнительные понятия: **технология** (совокупность методов и средств, направленных на достижение определенных целей); **инновационная деятельность** (процесс трансформации фундаментальных знаний в новые практические приложения [Рудь и др., 2011]), **технологическая инновация** (конечный результат инновационной деятельности, получивший воплощение в виде нового или усовершенствованного продукта или услуги, внедренных на рынке, нового или усовершенствованного технологического процесса или способа производства (передачи) услуг, используемых в практической деятельности) и некоторые другие понятия.

В качестве модели технологических трендов в данной работе используются результаты ОКР, выполняемой в НИУ ВШЭ по Госконтракту № 07.524.12.4018, где на основании выявленной системы базовых понятий была построена и реализована в системе Protégé [Protégé, 2012] онтологическая модель технологического тренда.

Вместе с тем, следует отметить, что в настоящее время модели технологических трендов, адекватные целям настоящей работы, которые бы акцептовались специалистами в данной области, отсутствуют. И более того, среди специалистов нет единого мнения о том, какие индикаторы определяют наличие и/или отсутствие описания тренда в коллекциях документов. Учитывая это, в настоящей работе предлагается к понятию технологического тренда «идти» через онтологическую модель выделения в документах системы терминов, которые потенциально могут представлять тренд на основе гибридного подхода.

1.3. Система автоматизации процессов выявления технологических трендов

1.3.1. *Общая архитектура системы АРМ «Тренд»*

Как показывает анализ литературы по методам и средствам автоматизированного выявления технологических трендов [Nallpati, 2003; Glance, et al., 2004; Yoon, et al., 2004; Shibata, et al., 2008; Kim, et al., 2009], в общей схеме обсуждаемой в настоящей работе АРМ «Тренд» целесообразно выделить следующие этапы:

- Выявление центров компетенции (включая организации и авторские коллективы) в области охвата исследуемого тренда.
- Формирование коллекций документов, соответствующих исследуемому тренду с учетом выявленных на предыдущем этапе центров компетенции.
- Собственно обработка сформированных коллекций документов.

Методы и средства автоматизированного формирования экспертных групп и выявления центров экспертной компетенции в определенных предметных областях представлены в работе [Хорошевский, 2010], формирование коллекций документов, соответствующих исследуемому тренду, является темой отдельной работы. Поэтому ниже обсуждаются вопросы собственно обработки сформированных коллекций документов.

При этом основными функциональными подсистемами в АРМ «Тренд» являются:

- Подсистема предварительной обработки сформированной коллекции документов.
- Подсистема гибридной обработки отдельных документов коллекции.
- Подсистема интеграции результатов обработки всей коллекции документов.
- Подсистема генерации онтологического представления результатов обработки всех коллекций документов.
- Подсистема визуализации и анализа результатов.

С учетом вышесказанного, общая архитектура АРМ «Тренд» может быть представлена схемой, показанной на Рисунок 1, а ее основные подсистемы обсуждается в следующем подразделе настоящей работы.

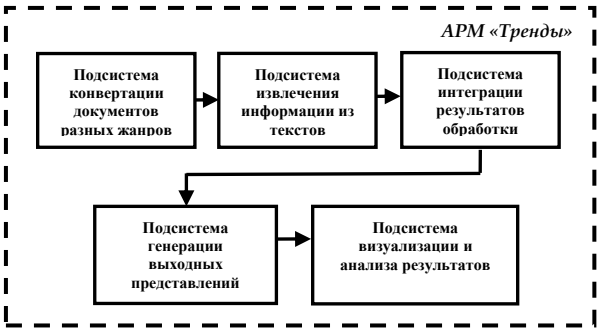

Рисунок 1 – Общая архитектура АРМ «Тренд»

1.3.2. *Функциональные подсистемы АРМ «Тренд»*

Подсистема предварительной обработки документов обеспечивает анализ форматов представления информации в сформированных коллекциях документов разных жанров и их планаризацию с помощью соответствующей системы конверторов.

Подсистема обработки отдельных документов коллекций на основе гибридного подхода предполагает использование лингвистических моделей извлечения информации из документов разных жанров и методов статистической обработки результатов извлечения информации из документов с помощью лингво-статистических процессоров, результатом работы которых являются характеристические вектора документов, соответствующие модифицированной модели «мешка слов» (Bag of Words).

В подсистеме интеграции результатов обработки всей коллекции документов осуществляется параметрическое слияние результатов обработки отдельных документов одной коллекции и результатов обработки разных коллекций.

Подсистема генерации выходных представлений результатов обработки всех коллекций документов поддерживает выбор выходного представления и собственно генерацию таких представлений в формате OWL и/или Tag Cloud.

Работа пользователя АРМ «Тренд» завершается в подсистеме визуализации и анализа результатов обработки коллекций документов.

Процессоры лингво-статистической обработки документов реализованы в АРМ «Тренд» с использованием инструментальной среды GATE [GATE, 2012], расширенной соответствующими модулями статистической обработки результатов извлечения терминов из текстов, а остальные подсистемы – в языке Java.

2. Выявление технологических трендов с использованием АРМ «Тренд»

2.1. Постановка задачи

Тестирование представленной выше АРМ «Тренд» осуществлялось в процессе обработки системы коллекций англоязычных документов, предоставленных ИСИЭЗ НИУ ВШЭ. Результат тестирования – информация о возможных технологических трендах, которые выявлены в процессе обработки исходных коллекций. Цель исследования – анализ полезности функционалов поддержки принятия решений экспертами-пользователями АРМ «Тренд» о присутствии в информационных материалах описаний технологических трендов.

2.2. Коллекции документов

Предоставленные для экспериментов исходные коллекции документов включали:

- аннотации научных статей (130867 док.),
- информацию из блогов (21 док.),
- информацию с новостных сайтов (560 док.),
- диссертации (29 док.) и некоторые другие коллекции.

Анализ предоставленных коллекций показал, что отдельные их документы, по сути дела, являются выгрузками из различных баз данных с различной структурой, не вполне соответствующих целям и задачам их последующей обработки. Поэтому для дальнейшего использования в АРМ «Тренд» было решено сконцентрироваться на коллекции аннотаций научных статей, предварительная планаризация которых была выполнена с помощью специально разработанного конвертора. В процессе предварительной обработки конвертор выделял в документе поля года публикации и аннотации публикации или, если таковое присутствует в документе, поле текста документа. Результатом работы конвертора было сохранение поля аннотации или текста публикации в отдельном планарном файле в коллекции, соответствующей году публикации. В случае отсутствия у документа поля года публикации, формировалась отдельная коллекция документов (Unknown) с неизвестным годом публикации.

2.3. Модуль гибридной обработки документов

Как показывает 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 системы терминов (однословных – Word и многословных – Expr).
- Предварительная статистическая обработка сформированной системы терминов.
- Формирование характеристических векторов документа (Bag of Words).
- Сохранение сформированных характеристических векторов в отдельных файлах.

В качестве инструментария для извлечения информации из текстов в настоящем исследовании использована платформа GATE, расширенная плагинами NP Chunker и Russian Morph Tagger [GATE, 2012], в котором используется открытая версия модуля русской морфологии компании Яндекс [Yandex, 2012].

Дополнительно к указанным выше компонентам извлечения информации из текстов для АРМ «Тренд» были разработаны специализированные ресурсы выделения простых именных групп в русскоязычных текстах, а также модуль генерации характеристических векторов документа (Bag of Words) и модуль сохранения сформированных характеристических векторов в отдельных файлах. Общие схемы цепочек ресурсов для обработки англоязычных и русскоязычных текстов представлены на рисунке 2.

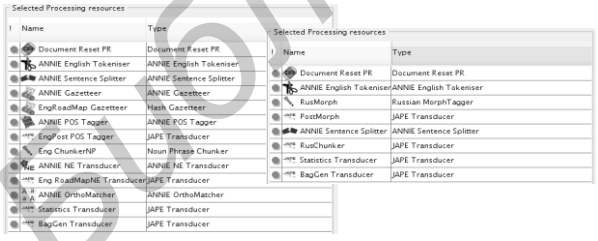

Рисунок 2 – Схемы цепочек ресурсов для обработки англоязычных и русскоязычных текстов

Для примера ниже показан фрагмент результатов работы модуля гибридной обработки англоязычного документа:

wos132472.txt
mesoporous_materials type=Expr;freq=1;TF=0.030303031
metal_oxides type=Expr;freq=1;TF=0.030303031
individual_polymer_chains type=Expr;freq=1;TF=0.030303031
fluorescent_dye type=Expr;freq=1;TF=0.030303031
oxide_nanostructures type=Expr;freq=1;TF=0.030303031
novel_nanocomposites type=Expr;freq=1;TF=0.030303031
molecular_imprinting type=Expr;freq=1;TF=0.030303031
shape_of_organic_molecules type=Expr;freq=1;TF=0.030303031
polyoxometalates type=Word;freq=1;TF=0.041666668
nanotechnology type=Word;freq=1;TF=0.041666668
……

Таким образом, на выходе модуля гибридной обработки документов формируются планарные файлы со следующей структурой:

- первая строка: имя файла (без расширения),
- каждая оставшаяся строка:
  - ключевое-слово | ключевое-выражение
  - type=Word | Expr,
  - freq=<целое> (количество повторений термина в документе),
  - TF=<десятичная-дробь> (частота встречаемости термина в документе).

2.4. Интеграция результатов обработки

Характеристические вектора, полученные в результате работы модуля гибридной обработки документов, обрабатываются специально разработанными и реализованными в рамках настоящего исследования модулями.

Первый модуль постобработки (ontoMerger) предназначен для интеграции характеристических векторов отдельных документов коллекции в единый характеристический вектор коллекции.

На вход данного модуля поступает коллекция характеристических векторов обработанных документов, а на выходе формируется единый файл характеристического вектора коллекции со следующей структурой:

- первая строка: N=<целое> (количество документов в коллекции),
- каждая оставшаяся строка:
  - ключевое-слово | ключевое-выражение
  - type=Word | Expr
  - DF=<целое> (количество документов в коллекции, в которых встречался термин);
  - TF=<десятичная-дробь> (средняя частота встречаемости термина в коллекции)

Второй модуль постобработки (ontoJointer) предназначен для объединения характеристических векторов отдельных коллекций в единый характеристический вектор системы коллекций. На вход данного модуля поступают характеристические вектора обработанных коллекций, а на выходе формируется единый файл характеристического вектора системы коллекций. При этом в процессе объединения характеристических векторов отдельных коллекций происходит фильтрация терминов по параметру DF, что позволяет установить порог отсечения терминов, входящих в общий характеристический вектор системы коллекций.

Последний модуль постобработки предназначен для расчета индексов терминов, входящих в систему коллекций. На вход этого модуля поступает файл

характеристического вектора системы коллекций, а на выходе формируется индексный файл терминов коллекции со следующей структурой:

- первая строка: N=<целое> (количество документов во всех коллекциях),
- каждая оставшаяся строка:
  - ключевое-слово | ключевое-выражение
  - type=Word | Expr
  - DF=<целое> (количество документов во всех коллекциях, где встречался термин);
  - TF=<десятичная-дробь> (средняя частота встречаемости термина в системе коллекций);
  - IDF=<десятичная-дробь> (обратная частота термина на системе коллекций)
  - TF*IDF=<десятичная-дробь> (индекс термина в системе коллекций).

Таким образом, в процессе постобработки документов отдельных коллекций и систем коллекций формируется информация, достаточная для последующего анализа экспертами на предмет наличия в системе коллекций статистически значимой информации о новых технологических трендах.

2.5. Генерация онтологического представления результатов

Для удобства анализа результатов обработки коллекций документов в рамках настоящего исследования был разработан и реализован специальный модуль (OWL Generator), который базируется на идеях, представленных в работе [Witte, et al., 2010]. На вход этого модуля поступает характеристический вектор коллекции (системы коллекций), а на выходе формируется ее OWL-представление, соответствующее онтологической модели тренда.

Полученные OWL-представления загружаются в систему онтологического инжиниринга Protégé и используются экспертами для дальнейшего анализа и принятия решения о наличии/отсутствии в обработанной системе коллекций документов информации о новых технологических трендах.

3. Предварительные результаты и направления дальнейших исследований и разработок

3.1. Коллекции документов

Для проверки базовых гипотез, сформулированных выше, были сформированы 11 коллекций научных публикаций за 2002-2012 г.г. общим объемом 130370 документов.

Каждая из коллекций была предварительно планаризована для дальнейшей обработки, каждый документ каждой коллекции был обработан с помощью реализованного гибридного модуля формирования характеристических векторов, а результаты были обработаны реализованными модулями интеграции результатов.

Полученные характеристические вектора поступали на вход реализованного модуля генерации OWL-представлений, на выходе которого формировалась экземплярная часть OWL-представления коллекции. Валидация полученных результатов осуществлялась в системе Protégé, которая в данном случае использовалась качестве инструментария онтологического инжиниринга.

3.2. Результаты обработки коллекций документов

После обработки всех коллекций были получены результаты, представленные в Таблица 1.

Таблица 1 – Статистика коллекций научных публикаций

Год	К-во док.	Терминов типа Word		Терминов типа Expr	
		Уник.	Всего	Уник.	Всего
2002	5762	4541	10729	13576	110785
2003	5916	4685	11065	13996	114314
2004	5885	4625	10898	13888	111987
2005	8976	6874	16018	28872	197315
2006	12136	8322	19058	38024	261994
2007	13038	8608	20243	39122	283060
2008	14031	8972	21047	40490	301267
2009	15181	9832	21290	59882	307716
2010	15748	9346	22682	40834	339256
2011	19053	11241	24700	71941	383765
2012	16145	12865	17926	140489	244760
Всего:	131871		184927		2545434

Для случая стандартных алгоритмов обработки документов всех коллекций распределение терминов представлено на рисунке 3, а для случая обработки с учетом только тех предложений, которые содержат индикаторы присутствия трендов в документах, - на Рисунок 4.

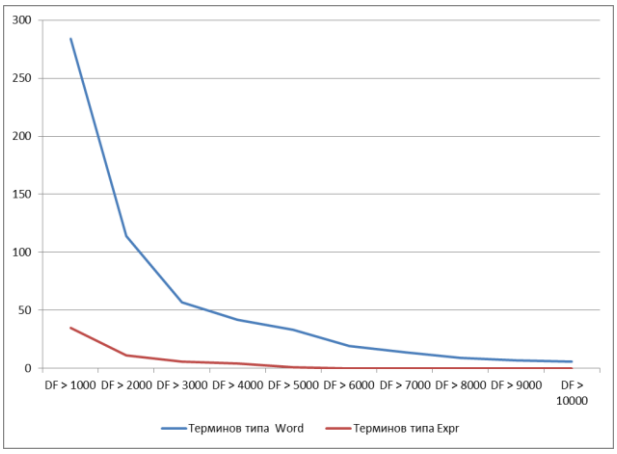

Рисунок 3 - Распределение терминов для случая стандартной обработки документов

Рисунок 4 - Распределение терминов для случая гибридной обработки документов с учетом индикативных предложений

Для примера, на Рисунок 5 приведено OWL-представление результатов обработки коллекций научных публикаций за 2002-2012 г.г. (DF>1500), а на Рисунок 6 - OWL-представление результатов обработки тех же коллекций для случая DF>2500.

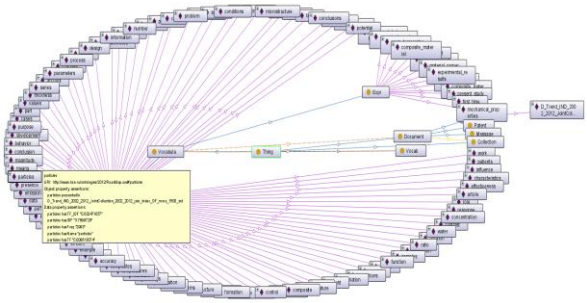

Рисунок 5 - OWL-представление коллекций научных публикаций за 2002-2012 гг. (DF>1500) в системе Protege

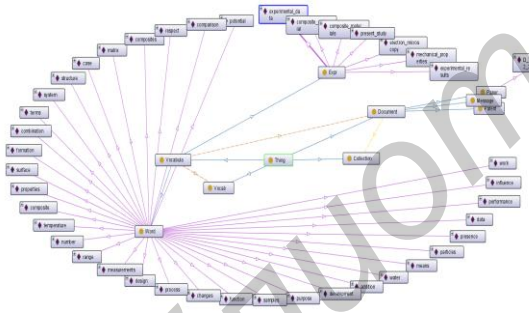

Рисунок 6 - OWL-представление коллекций научных публикаций за 2002-2012 гг. (DF>2500) в системе Protege

Как показывает анализ представленных выше результатов, в случае использования алгоритмов гибридной обработки документов с учетом только индикативных предложений количество выделенных терминов существенно меньше, чем в случае стандартных алгоритмов. Таким образом, использование алгоритмов, в которых обрабатывались только предложения, содержащие индикаторы возможного присутствия терминов трендов, позволяет существенно сократить объемы характеристических векторов коллекций, а динамика уменьшения числа выделенных терминов показывает более быстрое их сокращение. Вместе с тем, как показывает более детальный анализ результатов обработки одних и тех же коллекций научных публикаций, уменьшение числа обрабатываемых предложений в обеих схемах примерно одинаково. На наш взгляд, такая ситуация определяется тем, что коллекция научных публикаций в данном случае была представлена их аннотациями, в которых практически все предложения должны быть значимыми. В связи с этим представляет интерес дальнейший анализ коллекций полных текстов научных публикаций и сравнение результатов с результатами анализа аннотаций.

Особый интерес представляет анализ состава выделенных терминов в коллекциях научных публикаций. Для проведения такого анализа ниже представлены однословные (Таблица 2) и многословные (Таблица 3) термины для случая DF > 1000.

Таблица 2 – Термины типа Word (коллекции 2002-2012 г.г., DF > 1000, Количество документов в коллекции – 129833)

№/№	Key	DF	TF	IDF	TF*IDF
1.	ability	2253	0.005	4.053	0.023
2.	advantages	1048	0.003	4.819	0.016
3.	agreement	1034	0.003	4.832	0.016
4.	aircraft	1288	0.003	4.613	0.015
5.	carbon	1135	0.002	4.739	0.013
6.	composite	5169	0.016	3.223	0.052
…….					
	microstructure	2091	0.006	4.128	0.025
	nanocomposites	1316	0.003	4.591	0.016
	nanoparticles	1203	0.002	4.681	0.013
…….					
178.	water	3006	0.007	3.765	0.027

Таблица 3 – Термины типа Expr (коллекции 2002-2012 гг., DF > 1000, Количество документов в коллекции – 129833)

№/№	Key	DF	TF	IDF	TF*IDF
1.	carbon_nanotubes	1100	0.001	4.770	0.007
2.	composite_material	2728	0.005	3.862	0.020
3.	composite_materials	4842	0.008	3.288	0.029
4.	electron_microscopy	2557	0.003	3.927	0.014
5.	tensile_strength	1091	0.001	4.779	0.008
6.	thermal_conductivity	1151	0.001	4.725	0.008
7.	thermal_stability	1102	0.001	4.769	0.008
8.	thermogravimetric_analysis	1056	0.001	4.811	0.007
9.	transmission_electron_microscopy	1143	0.001	4.732	0.007
10.	x_ray_diffraction	2177	0.002	4.088	0.012

Как показывает анализ приведенных данных, среди терминов типа Word встречаются как интересные с точки зрения выявления новых

технологических трендов понятия, так и термины общей лексики.

Иная ситуация наблюдается в случае терминов типа Expr. Здесь практически все многословные термины представляют для экспертов интерес в плане выявления технологических трендов.

Дальнейший анализ полученных результатов осуществлялся путем построения временных рядов частоты появления многословных терминов в коллекциях документов разных лет по параметрам DF (Рисунок 7) и TF*IDF (Рисунок 8).

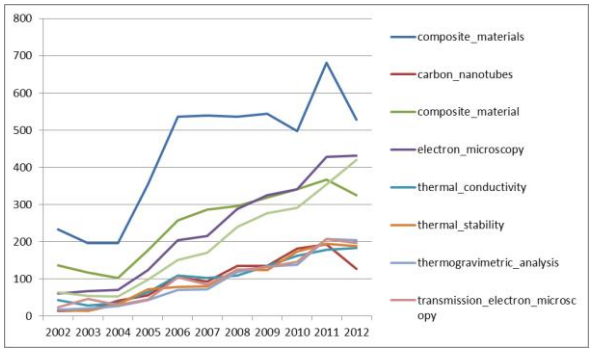

Рисунок 7 – Временной ряд частоты появления многословных терминов в коллекциях научных публикаций за 2002-2012 г.г. по параметрам DF

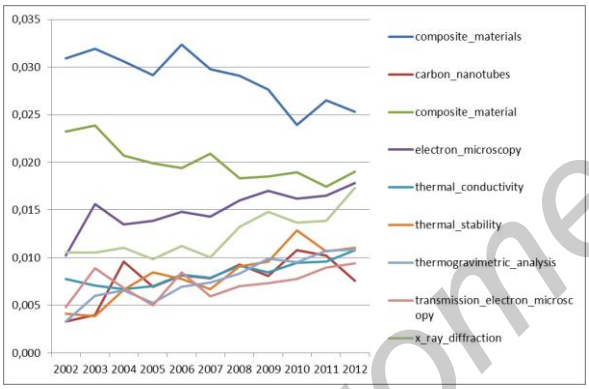

Рисунок 8 – Временной ряд частоты появления многословных терминов в коллекциях научных публикаций за 2002-2012 г.г. по параметрам TF*IDF

Как показывает анализ приведенных выше данных, по частоте встречаемости практически все многословные термины демонстрируют «бычий» тренд. В то же время, по параметру TF*IDF некоторые из многословных терминов демонстрируют «бычий» тренд, а другие – «медвежий» тренд.

Обсуждение полученных результатов с экспертами показало, что термины «бычьего» тренда могут идентифицировать ситуацию появления нового технологического тренда в области технологического триггера, а «медвежьего» – выход соответствующих результатов на плато продуктивности.

Таким образом, можно констатировать, что АРМ «Тренд» является полезным инструментом помощи экспертам в выявлении новых технологических трендов за счет автоматизации процессов обработки и анализа результатов на больших коллекциях документов.

Заключение

В работе рассмотрены вопросы разработки и реализации автоматизированного рабочего места поддержки процессов выявления новых технологических трендов АРМ «Тренд». Приведены результаты обработки представительной коллекции документов, в которых представлены аннотации научно-технических статей.

Предполагается, что направления дальнейших работ по автоматизации процессов выявления новых технологических трендов будут связаны с анализом групп индикаторов с точки зрения их эффективности и значимости для различных жанров документов и частотным анализом словосочетаний и шаблонов индикаторов; доработкой онтологических моделей трендов, задающих семантические поля индикаторов и стоп-слов; автоматизацией процессов анализа динамики изменения терминологического поля, описывающего технологический тренд; интеграцией программного обеспечения АРМ «Тренд» с инновационными средствами визуализации результатов анализа.

Благодарности. Работа выполнена при финансовой поддержке Минобрнауки России по государственному контракту от 16.05.2012 г. № 07.524.12.4018 в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007-2013 годы».

Библиографический список

[Рудь и др., 2011] Рудь В.А., Фурсов К.С. Роль статистики в дискуссии о научно-технологическом и инновационном развитии.// Вопросы экономики, 2011. № 1. С. 120—133.

[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

[GATE, 2012] Developing Language Processing Components with GATE Version 7 (a User Guide). http://gate.ac.uk/sale/tao

[Glance, et al., 2004] Natalie S. Glance, Matthew Hurst , Takashi Tomokiyo. BlogPulse: Automated trend discovery for weblogs // WWW 2004, Workshop on the webloging ecosystem: aggregation, analysis and dymanics, ACM, 2004.

[Kim, 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.

[Nallpati, 2003] R. Nallpati. Semantic language models for topic detection and tracking. In Proceedings of the conference of the North American chapter of the Association for Computational Linguistics on Human Language Technology (HLTNAACL' 03), 2003.

[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.

[Wiki, 2012] Понятие тренда. http://ru.wikipedia.org/wiki/Тренд

[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.

[Yandex, 2012] Морфология Яндекс. http://company.yandex.ru/technology/mystem/

[Yoon, et al., 2004] B. Yoon, and Y. Park. A text mining-based patent network: analytical tool for high-technology trend. Journal of High Technology Management Research, Vol. 15 (1), 2004.

TECHNOLOGY TRENDS WATCHING IN AUTOMATED WORKSTATION "TREND"

Khoroshevsky V. F.

Institution of Russian Academy of Sciences Dorodnicyn Computing Centre of RAS, Center for Information Intelligence Applications of Institute for Statistical Studies and Economics of Knowledge, NU HSE Moscow, Russia

khor@ccas.ru

vkhoroshevsky@hse.ru

Workstation ARM "Trend" oriented to the support of new technological trends watching based on a hybrid approach to information extracting is presented in the paper. The results of the processing of a representative collection of scientific and technical papers annotations are discussed.