УДК 004.62

Андрей Сергеевич Коляда
Аспирант кафедры управления системами безопасности жизнедеятельности

Одесский национальный политехнический университет, Одесса

ЛАТЕНТНО СЕМАНТИЧЕСКИЙ ПОДХОД ДЛЯ АНАЛИЗА ИНФОРМАЦИИ ИЗ НАУКОМЕТРИЧЕСКИХ БАЗ ДАННЫХ

Розглянуто особливості ідентифікації авторів та їх публікацій з наукометричних баз даних на основі латентно семантичного аналізу назв статей. Показано, що в слабоструктурованих наукових текстах назви статей одного автора утворюють загальну область термів, що дозволяє ідентифікувати авторів статей.

Ключові слова: публікації, бази даних, автори, ідентифікація, латентність даних, семантика, аналіз

Рассмотрены особенности идентификации авторов и их публикаций из наукометрических баз данных на основе латентно семантического анализа названий статей. Показано, что в слабоструктурированных научных текстах названия статей одного автора образуют общую область термов, что позволяет идентифицировать авторов статей.

Ключевые слова: публикации, базы данных, авторы, идентификация, латентность данных, семантика, анализ

Article covers features for authors and their publications identification in scientometric databases based on latent semantic analysis of article titles. It is shown that in poorly structured scientific texts titles of articles per author form a common area of terms that allows us to identify the authors of articles. Latent semantic analysis (LSA) uses a mathematical technique called singular value decomposition (SVD) to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text. It is based on the principle that words that are used in the same contexts tend to have similar meanings. A key feature of LSA is its ability to extract the conceptual content of a body of text by establishing associations between those terms that occur in similar contexts. LSA overcomes the following problems: multiple words that have similar meanings (synonymy) and words that have more than one meaning (polysemy). Another benefit of LSA is that it uses a strictly mathematical approach, so is inherently independent of language.

Keywords: publications, databases, authors, identification, data latency, semantics, analysis

Постановка проблемы

Проблема идентификации текстов естественного языка вычислительными машинами давно представляет научный интерес [1]. В настоящее время широко используются различные подходы распознавания речи, классификации текстовой информации, определения идентичности текстов [2]. Автоматизация извлечения информации из наукометрических баз данных сопряжена с необходимостью уточнения результатов запросов к базам данных в части исключения «двойников» авторов, у которых совпадают фамилии и инициалы [3 – 6]. Подобные проблемы возникают и при разработке систем автоматизированного обучения с открытыми тестами [7 – 10].

Одним из примеров извлечения смысла из текстов является поиск схожих документов или документов на определенную тематику. Стандартный поиск использует сравнение документов на наличие искомой строки или слов. Однако не всегда можно сформулировать точный запрос. Зачастую требуется поиск, который основан на анализе смысловой нагрузки документов. Одним из подходов, который уже активно используют поисковые гиганты, является латентно семантический анализ. Этот подход позволяет выявить закономерности в отношениях между понятиями и терминами в неструктурированной коллекции текстов.

Анализ последних исследований и публикаций

Существует несколько способов смыслового анализа текстов, которые можно разделить на следующие группы [11]: лингвистический анализ; статистический анализ.

Первая группа ориентирована на определении смысла по семантической структуре текста и включает лексический, морфологический и синтаксический анализ. В настоящее время отсутствуют сложившиеся подходы к реализации задачи семантического анализа текстовой информации, что во многом обусловлено исключительной сложностью проблемы и недостаточно полной проработкой научного направления создания систем искусственного интеллекта.

Вторая группа – это, как правило, частотный анализ в тех или иных его вариациях. Суть анализа заключается в подсчете количества повторений слов в тексте и использовании результатов подсчета для конкретных целей. Всевозможные варианты различных реализаций подсчета слов и последующая обработка результатов подсчета образуют широкий спектр предлагаемых в данном классе методов и алгоритмов.

Одним из наиболее эффективных статистических подходов является латентно семантический анализ (или латентно семантическое индексирование) [12]. Авторы представили модель двухрежимного факторного анализа, которая основана на сингулярном разложении (SVD). Сингулярное разложение представляет термины и документы в виде векторов в пространстве выбираемой размерности, а скалярное произведение между точками пространства – их схожесть.

Латентно семантический анализ начинается с построения матрицы документов и терминов – индексируемых слов [13]. Индексируемые слова это – слова, которые встречаются в двух или более документах и имеют смысловую нагрузку (не являются предлогами, союзами и т.д.). Далее применяется сингулярное разложение этой матрицы на произведение трех матриц:

$$A = U \cdot S \cdot V^t,$$

где матрицы U и V – ортогональные; S – диагональная матрица, диагональные значения которой называются сингулярными значениями матрицы A.

Такое разложение обладает замечательной особенностью: если в матрице S оставить только k наибольших сингулярных значений, а в матрицах U и V – только соответствующие этим значениям столбцы, то произведение получившихся матриц S, U и V будет наилучшим приближением исходной матрицы A к матрице $\hat{A}$ ранга k:

$$\hat{A} \approx A = U \cdot S \cdot V^t.$$

Основная идея латентно-семантического анализа состоит в том, что если в качестве матрицы A использовалась матрица, индексируемая слова на документах, то матрица Â, содержащая только k первых линейно независимых компонент A, отражает основную структуру различных зависимостей, присутствующих в исходной матрице. Структура зависимостей определяется весовыми функциями индексируемых слов. Таким образом, каждое индексируемое слово терм и документ представляются при помощи векторов в общем пространстве размерности k. Близость между любой комбинацией индексируемых слов и/или документов легко вычисляется при помощи скалярного произведения векторов [14]. Как правило, выбор зависит от поставленной задачи и подбирается эмпирически. Если выбранное значение слишком велико, то метод теряет свою мощность и приближается по характеристикам к стандартным векторным методам. Слишком маленькое значение k не позволяет улавливать различия между похожими термами или документами.

Латентно-семантический анализ полностью справляется с проблемой синонимии, но частично с проблемой полисемии потому, что каждое слово определяется одной точкой в пространстве. Также этот анализ позволяет выполнять автоматическую категоризацию документов, основанную на их сходстве концептуального содержания. Также преимуществом латентно семантического анализа является независимость от языка, так как это математический подход. Недостатком метода является снижение скорости вычисления при увеличении объема входных данных (например, при SVD-преобразовании).

Цель статьи

Извлеченная информация из наукометрических баз данных нуждается в постобработке с целью определения схожих по смыслу публикаций, а также определения дубликатов. Целью данной статьи является разработка способа семантического анализа извлеченной информации.

Основной материал исследования

Применение латентно семантического анализа для проекта по извлечению информации из наукометрических баз данных позволит разделить полученные публикации на категории с целью определения однофамильцев. Например, автор И.И. Иванов занимается исследованиями в области компьютерных наук, но результаты поиска его публикаций в наукометрических базах содержат много несоответствующих записей, так как есть еще один автор И.И. Иванов, который опубликовал статьи по медицинской тематике. Латентно семантический анализ позволит разделить

публикации, которые относятся к разным концептам. Различные наукометрические базы могут содержать дубликаты публикаций в несколько измененной форме. Определение этих дубликатов также возможно с помощью латентно семантического анализа.

Рассмотрим последовательность действий латентно семантического анализа, изображенную на рис. 1, к некоторому набору документов. В качестве примера будет использоваться небольшой список названий публикаций, извлеченных из наукометрических баз данных для автора Е.В. Колесникова (табл. 1). Выбор автора хорошо подходит для примера, так как существует несколько авторов с одинаковой фамилией и инициалами.

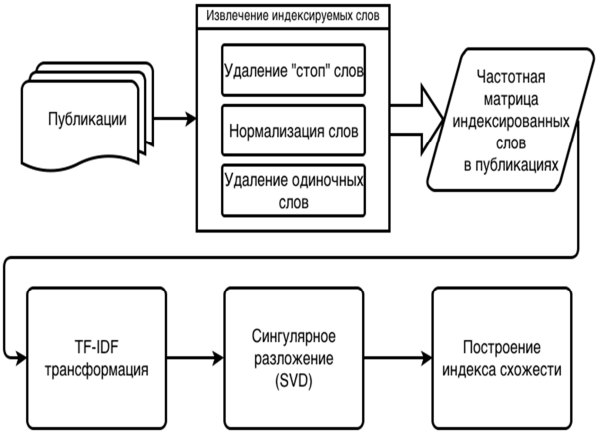

Рис. 1. Последовательность действий латентно семантического анализа

Таблица 1

Список документов для примера работы латентно-семантического анализа

Д1	Когнитивные модели слабо структурированных проектов создания программных продуктов
Д2	Лекарственно индуцированные поражения печени: особенности выявления, постановки диагноза и ведения пациентов
Д3	Трансформация когнитивных карт в модели марковских процессов для проектов создания программного обеспечения
Д4	Особенности поражения печени при ВИЧ инфекции
Д5	Матричная диаграмма и сильная связность индикаторов ценности в проектах
Д6	Патогенетические механизмы прогрессирования сочетанных вирусных и алкогольных поражений печени
Д7	Разработка марковских моделей изменений состояния пациентов в проектах предоставления медицинских услуг
Д8	Решенные и нерешенные вопросы терапии неалкогольной жировой болезни печени в рамках метаболического синдрома
Д9	Анализ структурной модели компетенций по управлению проектами национального стандарта Украины

Просмотрев список документов, можно заметить, что часть статей относится к медицинской тематике, а часть – к управлению проектами. Чтобы провести это разделение применяется латентно семантический анализ.

Изначально имеется список тем, который нужно проанализировать и обработать с целью выделения индексируемых слов.

Анализ включает в себя:

- удаление, так называемых, “стоп” слов, то есть, не имеющих смысловой нагрузки (предлоги, союзы и т.д.);

- приведение слов к нормальному виду или стемминг - процесс нахождения основы слова (используется алгоритм Портера [15], который позволяет быстро определить основу слова);

- удаление слов, встречающихся только один раз. Этот пункт не обязателен, но позволяет экономить ресурсы при расчетах.

На основе полученных индексируемых слов строится частотная матрица использования этих слов (табл. 2).

Таблица 2

Частотная матрица использования индексируемых слов в документах

Индексируемые слова	Документы								
	Д1	Д2	Д3	Д4	Д5	Д6	Д7	Д8	Д9
когнитивн	1		1						
марковск			1				1		
модел	1		1				1		
особен		1		1					
пациент		1					1		
печен				1		1		1	
поражен		1		1		1			
программн	1		1						
проект	1		1		1		1		1
создан	1		1						

Для повышения качества анализа используется следующий этап – трансформация матрицы с помощью модели TF-IDF (от англ. TF – term frequency, IDF – inverse document frequency) – статистическая мера, применяемая для оценки важности слова в контексте документа, являющегося частью коллекции документов. Вес некоторого слова пропорционален количеству употребления этого слова в документе и обратно пропорционален частоте употребления слова в других документах коллекции.

Следующий шаг, является основой латентно семантического анализа – это сингулярное разложение полученной матрицы и построение индекса схожести, который вычисляется по расстоянию между индексируемыми словами и документами в k-мерном пространстве. На рис. 2 показано графическое представление индексируемых слов и заголовков в двумерном пространстве (**k**=2), а таб. 3 содержит их координаты, полученные из сингулярного разложения.

Таблица 3

Координаты слов и документов в двумерном пространстве

Индексируемое слово	X	Y	Документ	X	Y
когнитивн	-0.33	0.05	**Д1**	-0.56	0.06
марковск	-0.28	-0.02	**Д2**	-0.04	-0.55
модел	-0.52	0.01	**Д3**	-0.64	0.05
особен	-0.01	-0.47	**Д4**	-0.01	-0.65
пациент	-0.12	-0.26	**Д5**	-0.15	0.00
печен	0.00	-0.52	**Д6**	-0.01	-0.47
поражен	-0.02	-0.66	**Д7**	-0.40	-0.10
программн	-0.33	0.05	**Д8**	0.00	-0.21
проект	-0.56	0.01	**Д9**	-0.29	0.01

Следует подчеркнуть, что разрешение указанных выше задач должно строиться на основе законов проектного управления [16]. При этом необходимо учитывать общие подходы к организации терминологических систем научного знания [17], модели коммуникационных процессов [18], а также особенности организации компьютерных сетей [19].

На рис. 2 показано, что часть документов принадлежит к одной тематике, а остальные – к другой. Проанализировав индексируемые слова, находящиеся рядом с документами, можно сделать вывод, что тематика 1 относится к управлению проектами (проект, модель, программа), а тематика 2 – к медицине (пациент, печен, поражен).

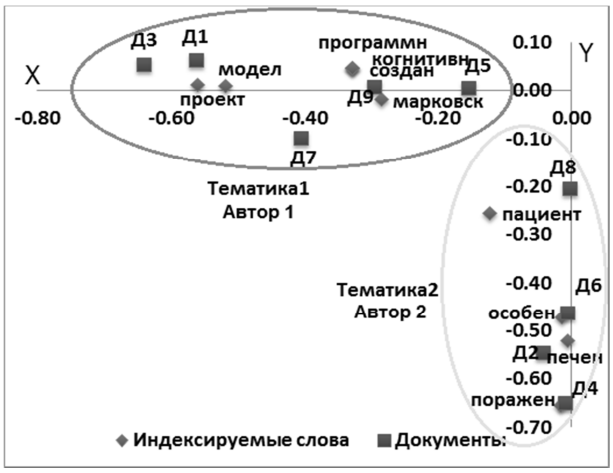

Рис. 2. Графическое представление распределения индексируемых слов и документов в двумерном пространстве

Выводы и перспективы дальнейших исследований

Латентно семантический анализ предоставляет довольно неплохие результаты сравнения различных документов по смыслу и дает возможность автоматической категоризации их. Будучи основанным на математических и статистических расчетах, этот подход является независимым от языка документов.

Применение латентно семантического анализа в проекте извлечения информации из наукометрических баз данных позволяет решать проблему однофамильцев и выявить дубликаты.

Список литературы

1. *Формализация проблемы извлечения знаний из естественно языковых текстов. / [А. Палагин, С. Кривый, Н. Петренко, Д. Бибиков]. – Sofia: Information technologies & knowledge, 2012. – 100 с.*
2. *Білощицький, А.О. Ефективність методів пошуку збігів у текстах / А.О. Білощицький, О.В. Діхтяренко // Управління розвитком складних систем. – 2013. – № 14.– С. 144 – 147.*
3. *Білощицький, А.О. Наукометричні бази та індикатори цитування наукових публікацій / А.О. Білощицький, В.Д. Гогунський // Інформаційні технології в освіті, науці та виробництві. – Вип. 4 (5). – О.: АО Бахва, 2013. – С. 198 – 203.*
4. *Коляда, А. С. Автоматизация извлечения информации из наукометрических баз данных / А. С. Коляда, В. Д. Гогунский// Управління розвитком складних систем. – 2013. – № 16. – С. 96 – 99.*
5. *Бурков, В. Н. Параметры цитируемости научных публикаций в наукометрических базах данных / В. Н. Бурков, А. А. Белощицкий, В. Д. Гогунский // Управління розвитком складних систем. – 2013. – № 15. – С. 134 – 139.*
6. *Коляда А. С. Разработка проекта информационно-аналитической системы извлечения и обработки информации из наукометрических баз данных / Коляда А. С., Негри А. А., Колесникова Е. В. // Управління проектами: стан та перспективи. Матеріали IX Міжнар. наук.-практ. конф. – Миколаїв: НУК, 2013. – 348 с.*
7. *Высоцкий, В.Ю. Разработка обучающих программ в виртуальной компьютерной среде // В.Ю. Высоцкий, В.Д. Гогунский // Тр. Одес. политехн. ун-та. – 2011. – Вып. 2 (36).– С. 184 – 189.*
8. *Яковенко, В.Д Комп'ютерна реалізація системи автоматизованого управління навчальним процесом // В. Д. Яковенко, В. Д. Гогунський, Г. Ф. Сафонова // Моделир. в прикладных научных исследованиях. Матер. XVI семинара. – Одеса: ОНПУ, 2008. – С. 27 – 30.*
9. *Тертышная, Т. И. Автоматизированная система контроля знаний / Т. И. Тертышная, Е. В. Колесникова, В. Д. Гогунский // Тр. Одес. политехн. ун-та. – 2001. – Вып. 1(13).– С. 125 – 128.*
10. *Яковенко, А.Е Стратегия принятия решений в условиях адаптивного обучения / А. Е. Яковенко, А. В. Нарожный, В. Д. Гогунский // Восточно-европейский журнал передовых технологий. – № 2/2 (14). – 2005. – С.105 – 110.*

11. Чугреев, В. Л. Модель структурного представления текстовой информации и метод ее тематического анализа на основе частотно-контекстной классификации / Санкт-Петербургский гос. электротехнический ун-т "ЛЭТИ" им. В.И. Ульянова. – 2003. – С. 25 – 29.

12. Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman (1990). Indexing by Latent Semantic Analysis. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE. 41(6). – С. 391-407.

13. Rehurek, R. (2011). Subspace tracking for latent semantic analysis. Advances in Information Retrieval. 289 – 300.

14. Roger B. Bradford (2008). An empirical study of required dimensionality for large-scale latent semantic indexing applications. In proceeding of: Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA.

15. Porter M.F. (1980). An algorithm for suffix stripping. Program, 14 n. 3, – pp. 130-137.

16. Гогунский, В.Д. Обоснование закона о конкурентных свойствах проектов / В. Д. Гогунский, С.В. Руденко, П.А. Тесленко // Управління розвитком складних систем. – 2012. – № 8.– С. 14 – 16.

17. Рач, В. А. Побудова термінологічної системи організації наукового знання [Текст] / В. Рач, О. Россошанська, О. Медведєва // Науковий світ. – 2011. – № 4. – С. 13 – 16.

18. Гогунський, В. Д. Марковські моделі комунікаційних процесів в міжнародних проектах / О. В. Власенко, В. В. Лебідь, В. Д. Гогунський // Управління розвитком складних систем. № 12. – 2012.– С. 35 – 39.

19. Плетнев, А.Н. Организация вычислительной сети студгородка «Политехник» с использованием оптического волокна / А.Н. Плетнев, А.Н. Миколюк, В.Д. Гогунский // Труды Одес. политехн. ун-та. – № 2(28). – Одеса: ОНПУ, 2007. – С. 138 – 140.

References

1. Palagin, A., Kriviy, S., Petrenko, N., Bibikov, D. (2012). Formalization of the problem of knowledge extraction from natural language texts. Information technologies & knowledge, 100 p.

2. Biloshchytskyi, A., Dihtyarenko, O. (2013). The effectiveness of methods for finding matches in texts. Management of complex systems, 14, pp. 144 – 147.

3. Biloshchytskyi, A., Gogunsky, V. (2013). Scientometric indicators and citation database of scientific publications. Information technology in education, science and production, 4, pp. 198 – 203.

4. Kolyada, A., Gogunsky, V. (2013). Automating the extraction of information from scientometric databases. Management of complex systems, 16.

5. Burkov, V., Beloschitsky, A. (2013). Options citation of scientific publications in scientometric databases. Management of complex systems, 15, p. 134 – 139.

6. Kolyada, A., Negri, A., Kolesnikova, E. (2013). Development of the information and analytical system for extraction and processing of scientometric databases. Project management: state and prospects. International scientific conference, 9, p. 348.

7. Visotskiy, V., Gogunsky, V. (2011). Development of educational software in a virtual computer environment. Proceedings of the Odessa National Polytechnic University, 2(36), pp. 184 – 189.

8. Yakovenko, V., Gogunsky, V., Safonova, G. (2008). Computer implementation of computer-aided learning management. Modeling in applied research, XVI seminar, pp. 27 – 30.

9. Ternishnaya, T., Kolesnikova, E., Gogunsky, V. (2001). Automated monitoring system knowledge. Proceedings of the Odessa National Polytechnic University, 1(13), pp. 125 – 128.

10. Yakovenko, A., Narozhny, A., Gogunsky, V. (2005). Strategy decisions under adaptive learning. East European Journal of Enterprise Technologies, 2/2(14), pp. 105 – 110.

11. Chugreev, V. (2003). Model structural representation of textual information and the method of its thematic analysis based on frequency content classification. St. Petersburg State Electrotechnical University "LETI", VI Ulyanov, pp. 25 – 29.

12. Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman (1990). Indexing by Latent Semantic Analysis. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE. 41(6), pp. 391 – 407.

13. Rehurek, R. (2011). Subspace tracking for latent semantic analysis. Advances in Information Retrieval, pp. 289 – 300.

14. Roger B. Bradford (2008). An empirical study of required dimensionality for large-scale latent semantic indexing applications. In proceeding of: Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA.

15. Porter M.F. (1980). An algorithm for suffix stripping. Program, 14 n. 3, – pp. 130 – 137.

16. Gogunsky, V., Rudenko, S., Teslenko, P. (2012). Justification law on competitive properties projects. Management of complex systems, 8, pp. 14 – 16.

17. Rach, V., Rossoshanska, O., Medvedeva, O. (2011). Building a terminological system of scientific knowledge. Science world, 4, pp. 13 – 16.

18. Gogunsky, V., Vlasenko, O., Lebid, D. (2012). Markov models of communication processes in international projects. Management of complex systems, 12, pp. 35 – 39.

19. Pletnev, A., Mikolyuk, A., Gogunsky, V. (2007). Organization campus computer network "Polytechnic" using an optical fiber. Proceedings of the Odessa National Polytechnic University, 2(28), pp. 138 – 140.

Статья поступила в редколлегию 22.01.2014

Рецензент: д-р техн. наук, проф. В.Д. Гогунский, Одесский национальный политехнический университет, Одесса.