Розглядається проблема порівняння документів за змістом на прикладі аналізу назв наукових публікацій для визначення близьких за тематикою. Для вирішення проблеми використовується латентно семантичний аналіз, застосування якого дозволяє отримати взаємозалежності між набором документів і словами, які вони містять. Також латентно семантичний аналіз надає можливість виділення ключових слів і класифікації документів за змістом

Ключові слова: ідентифікація, публікація, індексація, латентний, семантичний, аналіз, класифікація, інформація, сингулярний, матриця

Рассматривается проблема сравнения документов по смыслу на примере анализа названий научных публикаций для определения близких по тематике. Для решения проблемы используется латентно семантический анализ, применение которого позволяет получить взаимозависимости между набором документов и словами, которые они содержат. Также латентно семантический анализ предоставляет возможность выделения ключевых слов и классификации документов по смыслу

Ключевые слова: идентификация, публикация, индексация, латентный, семантический, анализ, классификация, информация, сингулярный, матрица

УДК 004.62

ДОСТОВЕРНОСТЬ ИДЕНТИФИКАЦИИ АВТОРСТВА НАУЧНЫХ ПУБЛИКАЦИЙ НА ОСНОВЕ ЛАТЕНТНО СЕМАНТИЧЕСКОГО АНАЛИЗА

А. С. Коляда
Аспирант*
E-mail: akolyada@gmail.com
В. Д. Гогунский
Доктор технических наук, профессор*
E-mail: vgog@i.ua
*Кафедра управления системами безопасности жизнедеятельности
Одесский национальный политехнический университет
пр. Шевченко, 1, г. Одесса, Украина, 65044

1. Введение

Латентно семантический анализ (ЛСА) – это техника анализа взаимозависимостей между набором документов и термами (словами), которые они содержат [1]. ЛСА предполагает, что близкие по смыслу термы встречаются в схожих частях текстов.

Практическими задачами с применением ЛСА являются:

– сравнение документов (их кластеризация и классификация);

– уахождение схожих документов на разных языках, после анализа базы переведенных документов;

– уахождение связей между термами (проблема синонимии и полисемии);

– уахождение документов по указанным термам.

Хотя техника не нова и уже используется многими компаниями (например, поисковые системы), число наглядных результатов ее работы достаточно мало.

2. Постановка проблемы и цель исследования

Одним из этапов извлечения и сбора информации является ее обработка. В отличие от технического процесса извлечения, обработка может представлять собой интеллектуальную и даже творческую работу. Основной задачей на этом этапе является определение достоверности результатов. Примером может служить следующая задача: заданы фамилия, имя и отчество (ФИО) автора и список публикаций, извлеченных по этим атрибутам; как определить статьи, только этого автора (так как атрибут запроса ФИО для разных авторов может совпадать.)

Цель данной работы заключается в разработке модели идентификации авторства научных публикаций с использованием дополнительной информации – достоверно известных названий некоторых статей данного автора.

Для достижения поставленной цели нужно решить следующие задачи:

– выделение ключевых слов из документов;

– определение схожести документов с заданными ключевыми словами.

3. Литературный обзор

Одним из основных способов извлечения знаний из текстовых коллекций является тематическое моделирование – способ построения модели коллекции текстовых документов, которая определяет, к каким темам относится каждый из документов. Модели со скрытыми (латентными) переменными оказались особенно эффективными для выявления скрытых структур в текстовых коллекциях. Предложено много мо-

делей для решения задач моделирования текстовых коллекций в таких приложениях, как классификация документов, поиск похожих документов, поиск экспертов, выявление сообществ и анализ временных трендов [2].

Одной из моделей тематического моделирования является латентно семантический анализ. Этот анализ начинается с построения матрицы документов и терминов – индексируемых слов [3]. 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. 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 [3]. Будучи основанным на математических и статистических расчетах, этот подход является независимым от языка документов. ЛСА используется в обработке естественного языка, когнитивной науке и компьютерной лингвистике. Результаты применения ЛСА можно показать графически с представлением отношения термов и документов в двумерном пространстве [4], что позволяет наглядно увидеть взаимозависимости документов и термов (рис. 1).

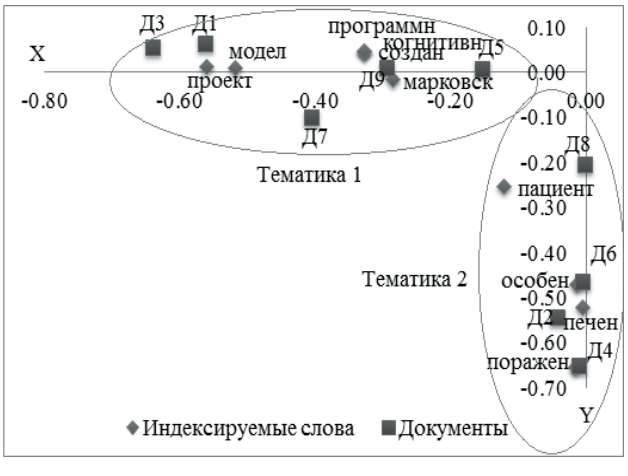

Рис. 1. Графическое представление распределения индексируемых слов и документов в двумерном пространстве

Для повышения качества анализа можно использовать метод взвешенных дескрипторов [5], в котором анализ встречающихся в тексте слова производится не только по частоте, но и учитывая семантику, за счет подбора соответствующих дескрипторов. Для получения значимых концептуальных дескрипторов используют законы Джорджа Зипфа, который предложил, что слова с большим количеством букв встречаются реже коротких слов.

Наиболее известными практическими реализациями ЛСА на сегодняшний день являются:

– *Sense Clusters*. Основная функция – кластеризация схожих контекстов. Применяется при разрешении неоднозначности слов (в частности, имен), классификации документов разного рода (электронных писем, новостных статей), классификация лексики (нахождение синонимов, антонимов и других классов отношений) [6].

– *S-Space*. Основная функция – универсальное средство для построения и обработки векторной модели. Содержит реализации большого количества алгоритмов (разные векторные модели, некоторые методы их последующей обработки). Ориентировано на скорость работы, интуитивно понятное представление данных [7].

– *Gensim*. Наиболее надежное и эффективное программное обеспечение, которое реализует семантическое моделирование для обычного текста. Предназначено специально для обработки больших коллекций документов, с использованием эффективных алгоритмов [8].

Дальнейшим развитием латентно семантического анализа является вероятностный латентно семантический анализ (ВЛСА) [9] – статистическая модель анализа автоматизированной индексации документов. Здесь, также как и ЛСА, каждый документ представляется числовым вектором, каждая компонента которого равна доле соответствующей темы в документе, но весовые коэффициенты слов определяются с помощью вероятностной модели. Для построения вероятностной модели можно использовать EM-алгоритм (Expectation-Maximization) – алгоритм, используемый в математической статистике для нахождения оценок максимального правдоподобия параметров вероятностных моделей, в случае, когда модель зависит от некоторых скрытых переменных. Однако вероятностная модель не описывает ни закон распределения этих долей, ни вероятности самих документов. В результате число параметров модели линейно растёт с ростом размера текстовой коллекции, что может приводить к переобучению.

4. Модель определения авторства научных публикаций

Проект по извлечению информации из наукометрических баз данных (НМБД) подразумевает получение информации о публикациях, которые принадлежат конкретному автору, из наиболее известных НМБД [10]. Выполнение поиска по заданному аргументу – ФИО – позволяет получить список публикаций, автором которых, по идее, является один человек. Но это не всегда верно, так как ФИО автора не может быть уникальным идентификатором записи. В мире могут существовать несколько авторов с одинаковыми ФИО. Добавим к этому тот факт, что чаще всего публикации содержат только инициалы с фамилией, поэтому вероятность нахождения публикаций нескольких авторов с идентичными ФИО, еще выше. Поэтому для выборки публикаций принадлежащих одному автору, нужно использовать дополнительную информацию из доступных полей структуры данных – название публикации. Название может отражать направление деятельности автора, а также это обязательное поле, которое не может быть пустым, в то время как остальные поля зачастую не доступны в тех или иных наукометрических базах [11]. На рис. 2 показан алгоритм использования ЛСА для классификации публикаций.

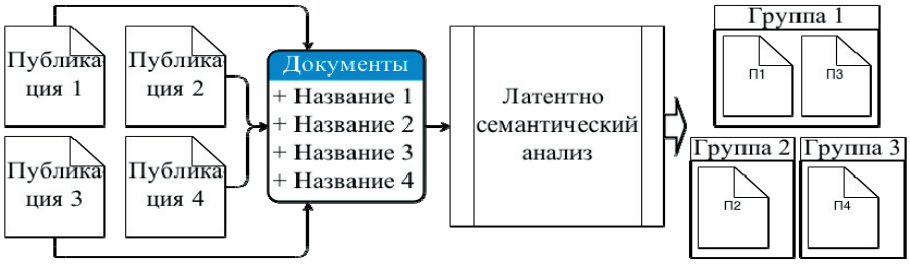

Рис. 2. Применение латентно семантического анализа для определения авторства научных публикаций

4. 1. Применение латентно семантического анализа к набору публикаций

Рассмотрим применение латентно семантического анализа при анализе названий публикаций. В табл. 1 представлен пример результатов поиска публикаций для автора “Колесникова Е. В.”.

Таблица 1

Фрагмент публикаций извлеченных по запросу “Колесникова Е. В.”

№	НМБД	Название публикации
1	Base-search	Лекарственно-индуцированные поражения печени: особенности выявления, постановки диагноза и ведения пациентов
2	Base-search	Современное состояние проблемы самоубийств в судебной медицине
3	Base-search	К вопросу о патоморфологических исследованиях нейроэндокринной системы при завершенных суицидах
4	Base-search	Теоретические исследования рабочего цикла гидравлического устройства ударного типа для ликвидации прихватов бурового снаряда в разведочных скважинах
5	Base-search	Гипоадипонектимия - ключевой фактор риска неалкогольной жировой болезни печени (обзор литературы)
6	Base-search	Особенности диагностики при подозрении на диффузную форму рака молочной железы
7*	Base-search	Трансформация когнитивных карт в модели марковских процессов для проектов создания программного обеспечения
8*	Base-search	Развитие теории проектного управления: обоснование закона К. В. Кошкина о завершении проектов

** - публикации искомого автора “Колесникова Е. В.”*

Как видно из приведенных данных, часть статей связана с медицинской тематикой, но последние две публикации (отмеченные звездочками) относятся к совершенно иной предметной области. Если известно направление деятельности автора, то можно определить с некоторой погрешностью, какие из публикаций принадлежат данному автору. Для того чтобы этот процесс автоматизировать, можно выделить ключевые слова из сферы предметной области автора и с помощью программы отобрать подходящие варианты.

Но тут возникает проблема: нам нужен набор из множества слов, которые могут встречаться в названиях статей. Этот набор может быть слишком объёмным, что скажется на производительности. Кроме этого следует принимать во внимание, что некоторые слова могут употребляться в разных контекстах с разным смыслом (проблема полисемии).

Чтобы решить эти проблемы используется латентно семантический анализ, который позволяет выделить связь между ключевыми словами и набором документов (названий публикаций, в нашем случае). Допустим, задано ключевое слово “информация”.

Применение латентно семантического анализа позволяет установить скрытые связи, например, слова “программа” или “компьютер” близки к предметной области этого слова. Поэтому ЛСА позволяет получить не только список публикаций, где встречается слово “информация”, но и без этого слова с наиболее близкими по смыслу [12].

Применение ЛСА можно разбить на несколько этапов (рис. 3):

– извлечение индексируемых слов: удаление слов, не имеющих смысловой нагрузки; удаление слов, встречающихся только один раз во всех документах; нормализация слов (выделение основы слов);

– построение частотной матрицы использования индексируемых слов в документах;

– TF-IDF трансформация частотной матрицы – статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью списка анализируемых документов;

– сингулярное разложение полученной матрицы, что позволяет отразить основную структуру различных зависимостей, присутствующих в исходной матрице. Каждый терм и документ представляются при помощи векторов в общем пространстве размерности k – количество наибольших сингулярных значений;

– построение индекса схожести – расчет близости между заданными ключевыми словами и документами. На практике, рассчитывается косинус угла между векторами (1).

$$\cos\theta = \frac{d \cdot q}{\|d\|\|q\|}, \quad (1)$$

где d – вектор документа, q – вектор ключевых слов, d·q – скалярное произведение векторов, $\|d\|$ и $\|q\|$ – норма векторов, которые рассчитываются по формуле (2):

$$\|q\| = \sqrt{\sum_{i=1}^{n} q_i^2}. \quad (2)$$

Значения косинуса угла между векторами, близкие к 1 представляют собой очень похожие слова, в то время как значения, близкие к 0 представляют собой очень разнородные слова.

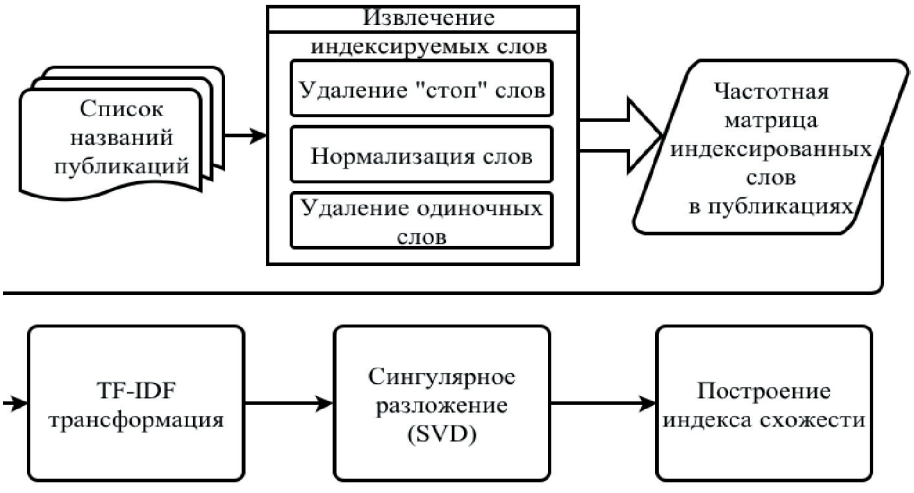

Рис. 3. Последовательность действий латентно семантического анализа

Таблица 2

Результат латентно семантического анализа по заданным ключевым словам

%	Наукометрическая база	Название публикации
88.75	Science Index	Матричная диаграмма и «сильная связность» индикаторов ценности в проектах
85.78	Science Index	Методы оценки проектов и программ
76.52	Base-search	Разработка марковских моделей изменений состояния пациентов в проектах предоставления медицинских услуг
69.47	Base-search	Трансформация когнитивных карт в модели марковских процессов для проектов создания программного обеспечения
55.57	Copernicus	Управление знаниями в IT-проектах
51.27	Base-search	Составляющие поведенческой компетенции участника команды проекта на основе компетентностного подхода
47.99	Base-search	Анализ структурной модели компетенций по управлению проектами национального стандарта Украины
...		
-0.37	Base-search	К вопросу о патоморфологических исследованиях нейроэндокринной системы при завершенных суицидах

По этим результатам можно предположить, что публикации с высоким уровнем схожести принадлежат нашему автору, а с низким – иным авторам.

4. 2. Классификация публикаций и извлечение ключевых слов

Еще одним вариантом применения латентно семантического анализа является разбиение документов на некоторые группы, связанные по смыслу. Так как ЛСА подразумевает наличие скрытых тем, к которым относятся термины и документы, можно выделить ключевые слова заданного количества тем. Например, мы не можем выделить ключевые слова для направления научной деятельности. С помощью латентно семантического анализа можно проанализировать список названий публикаций и разбить их, допустим, на 3 части и предоставить для каждой группы набор ключевых слов (табл. 3).

По этому списку ключевых слов можно установить, что третья группа относится к тематике нашего автора. Выполнив поиск по ключевым словам этой группы, можно получить основную часть статей нашего автора, без публикаций медицинской тематики. Точнее, более высокий уровень схожести будет получен для публикаций нашего автора [13–15]. Можно отбрасывать публикации, уровень схожести которых не превышает заранее заданный порог.

Таблица 3

Ключевые слова, соответствующие смысловым группам

№ группы	Ключевые слова
1	печен, неалкогольн, жиров, болезн
2	систем, исследован, суицид, патоморфологическ
3	проект, процесс, управлен, состоян

Ключевые слова, предложенные латентно семантическим анализом, можно сохранить и в следующий раз использовать их при новом поиске публикаций этого автора. Таким образом, можно создать обучаемую систему в полуручном режиме и использовать в автоматическом. Результат латентно семантического анализа, конечно же, может иметь погрешность. Это хорошо заметно, когда ключевые слова можно отнести к различным предметным областям научной деятельности. Например, слово “проект” может использоваться в любой сфере: учебный проект, медицинский проект, управление проектами и т. д. При этом в документах с малым количеством слов, ключевые слова могут иметь большую весомость.

5. Выводы

Применение латентно семантического анализа позволяет выделить ключевые слова для скрытых тем, на основе которых можно увидеть разброс тематик заданных публикаций. Для определения публикаций некоторого автора, рассчитывается коэффициент схожести заданных ключевых слов с набором публикаций. Это позволяет в полуавтоматическом режиме выделять ключевые слова тематик публикаций автора и автоматизировать отброс публикаций, несоответствующих этим темам. Перспектива дальнейших исследований состоит в определении погрешности результатов латентно семантического анализа и минимизации этой погрешности.

Литература

1. Deerwester, Scott Indexing by Latent Semantic Analysis [Text] / Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman // Journal of the American society for information science. – 1990. - № 41(6). – P. 391-407.
2. Daud, Ali Knowledge discovery through directed probabilistic topic models: a survey [Text] / Ali Daud, Juanzi Li, Lizhu Zhou, Faqir Muhammad // In Proceedings of Frontiers of Computer Science in China, 2010. – P. 280–301.
3. Řehůřek, R. Subspace tracking for latent semantic analysis [Text] / R. Řehůřek. – Advances in Information Retrieval, 2011. - P. 289–300.
4. Коляда, А. С. Латентно семантический подход для анализа информации из наукометрических баз данных [Текст] / А. С. Коляда // Управління розвитком складних систем. – 2014. – Вып. 17. – С. 90–94.
5. Стенин, А. А. Латентно-семантический метод извлечения информации из интернет ресурсов [Текст] / А. А. Стенин, Ю. А. Тимошин, Е. Ю. Мелкумян, В. В. Курбанов // Восточно-Европейский журнал передовых технологий. – 2013. – Т. 4, № 9 (64). – С. 19–22.
6. Pedersen, T. Duluth: Word Sense Induction Applied to Web Page Clustering [Text] : proc. of the 7th inter. workshop / T. Pedersen // Semantic Evaluation (SemEval 2013), in conjunction with the Second Joint Conference on Lexical and Computational Semantics (*SEM-2013), 2013. – P. 202–206.
7. Jurgens, D. The S-Space Package: An Open Source Package for Word Space Models [Text] : proc. ACLDemos ‘10 / D. Jurgens, K. Stevens // Proceedings of the ACL System Demonstrations, 2010. – P. 30–35.
8. Řehůřek, R. Software Framework for Topic Modelling with Large Corpora [Text] : proc. of the LREC 2010 workshop / R. Řehůřek, P. Sojka // New Challenges for NLP Frameworks, 2010. – P. 45–50.
9. Hofmann, T. Probabilistic Latent Semantic Indexing [Text] : proc. of the twenty-second annual inter. SIGIR conf. / T. Hofmann // Research and Development in Information Retrieval, 1999. – P. 50–57.
10. Коляда, А. С. Управління проектами: стан та перспективи [Текст] : матеріали IX міжнар. наук.-практ. конф. / А. С. Коляда, А. А. Негри, Е. В. Колесникова. – Миколаїв : НУК, 2013. – 348 с.
11. Коляда, А. С. Автоматизация извлечения информации из наукометрических баз данных [Текст] / А. С. Коляда, В. Д. Гогунский // Управління розвитком складних систем. – 2013. – № 16. – С. 96–99.
12. Roger, B. Bradford An empirical study of required dimensionality for large-scale latent semantic indexing applications [Text] : proc. of the 17th ACM conf. / B. Roger Bradford // Information and Knowledge Management, 2008. – P. 153–162.
13. Палагин, А. Формализация проблемы извлечения знаний из естественно языковых текстов [Текст] / А. Палагин, С. Кривый, Н. Петренко, Д. Бибиков // Information technologies & knowledge, 2012. – 100 с.
14. Бурков, В. Н. Параметры цитируемости научных публикаций в наукометрических базах данных [Текст] / В. Н. Бурков, А. А. Белощицкий, В. Д. Гогунский // Управління розвитком складних систем. – 2013. – № 15. – С. 134–139.
15. Білощицький, А. О. Наукометричні бази та індикатори цитування наукових публікацій [Текст] / А. О. Білощицький, В. Д. Гогунський // Інформаційні технології в освіті, науці та виробництві. – 2013. – Вип. 4 (5). – С. 198–203.