ISSN 2076-2429 (print)
ISSN 2223-3814 (on line)

А.В. Соколов. **Быстродействующий генератор ключевых последовательностей на основе клеточных автоматов.** Одним из наиболее эффективных современных подходов к решению задачи потокового шифрования является подход, основанный на использовании для генерации псевдослучайных ключевых последовательностей математического аппарата булевых функций. Предлагается новая конструкция генератора псевдослучайных ключевых последовательностей на основе клеточных автоматов и таких совершенных алгебраических конструкций, как бент-функции и последовательности со свойством *k*-граммного распределения. Установлено, что генератор генерирует высококачественные с точки зрения нелинейности и стохастических свойств ключевые последовательности, при этом по сравнению с существующими имеет простую программную реализацию. Таким образом, предложенный генератор псевдослучайных ключевых последовательностей может быть рекомендован для использования в современных криптографических приложениях, например, алгоритмах поточного шифрования, требующих высококачественные псевдослучайные ключевые последовательности.
Ключевые слова: ключевая последовательность, клеточный автомат, бент-функция, последовательность со свойством *k*-граммного распределения.

A.V. Sokolov. **Quick key sequences generator based on cellular automata.** One of the most effective modern approaches to the problem of stream encryption used for generation of pseudo-random key sequences is based on the Boolean functions mathematical instrument. A new design of the pseudo-random key sequences generator based on cellular automata and such perfect algebraic structures as bent functions and sequences with the property of *k*-gram distribution is proposed. It is shown that new generator generates a highly nonlinear and good stochastic quality pseudo-random key sequences compared with the existing high-speed generators, and have a simple program implementation. Thus, the proposed pseudo-random key sequences generator can be recommended for use in modern cryptographic applications, for example, stream encryption algorithms which require high-quality pseudo-random key sequences.
Keywords: key sequence, cellular automata, bent function, sequence with the property of *k*-gram distribution.

Рецензент д-р техн. наук, проф. Одес. нац. политехн. ун-та Мазурков М.И.

Поступила в редакцию 21 февраля 2014 г.

УДК 004.62

А.С. Коляда, магистр,
В.А Яковенко, магистр,
В.Д. Гогунский, д-р техн. наук, проф.,
Одес. нац. политехн. ун-т

ПРИМЕНЕНИЕ ЛАТЕНТНОГО РАЗМЕЩЕНИЯ ДИРИХЛЕ ДЛЯ АНАЛИЗА ПУБЛИКАЦИЙ ИЗ НАУКОМЕТРИЧЕСКИХ БАЗ ДАННЫХ

Введение. Проект по извлечению информации из наукометрических баз данных (НМБД) [1] подразумевает получение информации о публикациях из наиболее известных НМБД, которые принадлежат конкретному автору. Так как в мире может существовать несколько людей с одинаковыми ФИО, это поле не может быть уникальным идентификатором записи. Добавив к этому тот факт, что чаще всего публикации содержат только инициалы с фамилией, вероятность нахождения публикаций нескольких авторов с идентичными ФИО, еще выше. Для решения этой проблемы используется латентно-семантический анализ (ЛСА) [2], который позволяет вы-

DOI: 10.15276/opu.1.43.2014.32

делить семантическую связь между названиями публикаций и по заданным ключевым словам отбросить нерелевантные публикации.

Одним из недостатков ЛСА является то, что вероятностная модель метода не соответствует реальности. Предполагается, что слова и документы имеют нормальное распределение, хотя более приближенным к реальности является распределение Пуассона. Также наблюдается значительное снижение скорости вычисления при увеличении объема входных данных. Появляется проблема минимизации недостатков ЛСА и адаптации ее к задаче определения авторства научных публикаций.

Анализ последних исследований и публикаций. 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 [3]. Вероятностная тематическая модель (ВТМ) описывает каждую тему дискретным распределением на множестве терминов, а каждый документ — дискретным распределением на множестве тем. Предполагается, что коллекция документов — это последовательность терминов, выбранных случайно и независимо из множества таких распределений. Задачей тематического моделирования является восстановление компонент множества по выборке. Так как документ или термин может относиться одновременно ко многим темам с различными вероятностями, ВТМ осуществляет так называемую “мягкую” их кластеризацию. Это позволяет решать проблемы синонимии и омонимии терминов, возникающие при “жёсткой” кластеризации (документ или термин относится только к определенной тематике).

Модели со скрытыми (латентными) переменными оказались особенно эффективными для выявления скрытых структур в текстовых коллекциях [4], которые используются для решения таких задач, как классификация документов, поиск схожих документов, многоязычный поиск, выявление ключевых слов в документе, выявление зависимостей между терминами, выявление трендов в различных областях интересов и др.

Базовыми тематическими моделями являются:

— Вероятностный латентный семантический анализ (probabilistic latent semantic analysis, PLSA) — это статистическая модель анализа автоматизированной индексации документов (рис. 1). Данная модель является дальнейшим развитием латентно семантического анализа и основана на введении скрытых переменных — тематик текстовых документов. Хотя эта модель и считается улучшением латентно-семантического анализа, все таки она имеет существенные недостатки. Для PLSA характерно переобучение, а также неоднозначноть результатов, связанные с большим количеством вероятностных параметров, на которые не накладываются ограничения регуляризации. Также модель не выделяет нетематические слова. Большинство из недостатков может быть устранено применением семплирования, регуляризации и разреживания, что приводит к созданию большого семейства алгоритмов на основе PLSA.

— Модель латентного размещения Дирихле (latent Dirichlet allocation, LDA) — порождающая модель, в которой каждый документ рассматривается как смесь различных тем (рис. 2). Эта модель схожа с PLSA, но отличается тем, что в LDA распределение тем следует распределению Дирихле. Это позволяет оценивать вероятности документов и терминов вне текстовой коллекции. Для идентификации параметров модели LDA по коллекции документов применяется семплирование по Гиббсу или вариационный байесовский вывод. Эта модель устраняет основные недостатки PLSA, в частности, число параметров не увеличивается с ростом числа документов. Основной недостаток латентного распределения Дирихле — отсутствие лингвистических обоснований, хотя существуют ее расширения, которые устраняют некоторые её ограничения и повышают производительность для конкретных задач.

Из-за большого количества приложений и обобщений, модель латентного размещения Дирихле лидирует среди вероятностных тематических моделей.

Примеры расширений LDA:

— Автор-тематическая модель (author-topic model), которая представляет собой расширение LDA для выявления зависимостей между документами и авторами, выявление интересов автора.

ISSN 2076-2429 (print)
ISSN 2223-3814 (on line)

— Скрытая тематическая модель гипертекста (latent topic hypertext model, LTHM) использует связи между словами и ссылками для более точного определения тематики текста.

— Композитная модель HMM-LDA представляет зависимости между синтаксисом и семантикой текста, состоит из скрытой марковской модели (HMM) и LDA. Первая описывает закономерности между соседними словами, а вторая — глобальное тематическое описание документа в целом.

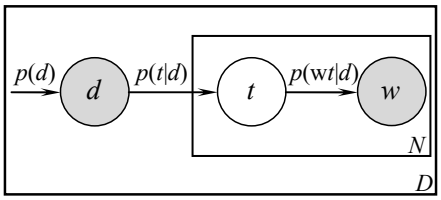

Рис. 1. Графическая вероятностная модель PLSA: d — документ; w — слово; d, w — наблюдаемые переменные; t — тема (скрытая переменная); p(d) — априорное распределение на множестве документов; p(w|t), p(t|d) — искомые условные распределения; D — коллекция документов; N — длинна документа в словах

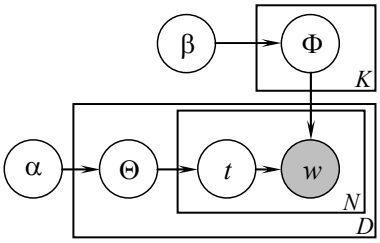

Рис. 2. Графическая вероятностная модель LDA: w — слово (наблюдаемая переменная); t — тема (скрытая переменная); D — коллекция документов; N — длинна документа в словах; K — количество тем в коллекции; Θ — распределение тем в документе; Φ — распределение слов в теме; α — априорное распределение Дирихле на параметры Θ, β — априорное распределение Дирихле на параметры Φ

Цель работы заключается в определении наиболее подходящей тематической модели для классификации научных публикаций по авторам-однофамильцам.

Изложение основного материала. Исходя из анализа существующих вероятностных тематических моделей, модель латентного размещения Дирихле [5] является хорошим кандидатом на использование в проекте по извлечению публикаций из наукометрических баз данных. После извлечения публикаций из наукометрических баз данных получаем список их названий. Задачей LDA является автоматическое определение тем, которые содержат эти названия. Применение LDA дает модель (рис. 3).

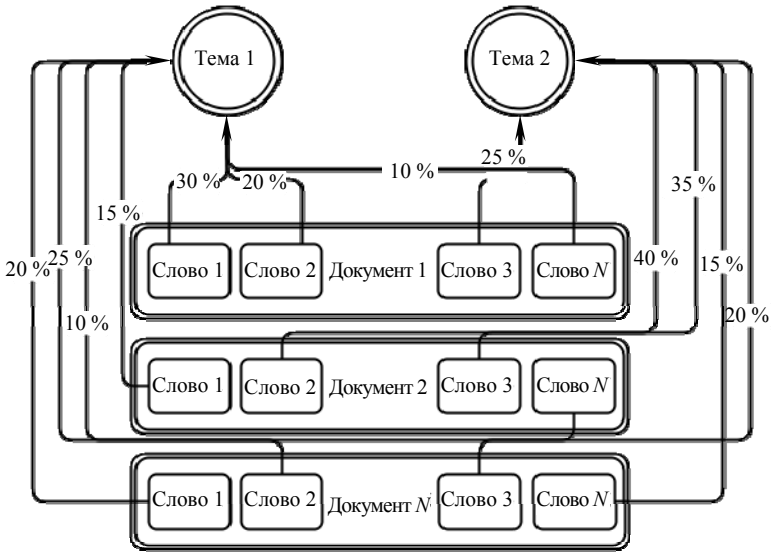

Рис. 3. Модель LDA с выделением двух тем

На рис. 3 показаны распределения слов в двух темах. Исходя из этого, можно судить, что документ 1 больше относится к первой теме, чем ко второй (30 % + 20 % + 10 % против 25 %), а также получить список наиболее подходящих к теме слов.

Для идентификации параметров модели LDA по коллекции документов можно применить семплирование по Гиббсу — алгоритм для генерации выборки совместного распределения множества случайных величин. Он используется для оценки совместного распределения и для вычисления интегралов методом Монте-Карло. Допустим, необходимо определить K тем в наборе документов, тогда алгоритм семплирования по Гиббсу можно описать так:

— Для каждого слова из каждого документа присвоить случайным образом одну тему (t) из K возможных;

— Для каждого слова из каждого документа вычислить:

– $p(t|d)$ — пропорция слов в документе d, которые присвоены теме t;

– $p(w|t)$ — пропорция слова w во всех документах, присвоенного к теме t;

– присвоить слову w новую тему t с вероятностью $p(t|d)\times p(w|t)$.

— Повторить второй пункт несколько раз (количество итераций также является входным параметром).

Метод LDA основан на вероятностной модели

$$p(d,w)=\sum_{t\in T} p(d)\cdot p(w\,|\,t)\cdot p(t\,|\,d),$$

где d — документ;

t — тема;

w — слово;

T — множество тем;

$p(d)$ — априорное распределение на множестве документов;

$p(w|t)$ — условное распределение слова w в теме t;

$p(t|d)$ — условное распределение темы t в документе d.

Результаты. Для демонстрации результатов анализа возьмем набор публикаций, полученный по параметру поиска “Яковенко В.Д.”, и выполним сравнение с ключевыми словами “система”, “автоматизований”. Процедура выполняется также с использованием ЛСА с двумя повторениями (см. таблицу).

Результат сравнения публикаций с ключевыми словами

ЛСА, %		LDA, %		Публикация
1	2	1	2	
82	82	100	100	Прогнозирование состояния системы управления качеством деятельности учебного заведения
0	0	52	53	К вопросу о причинно-следственных взаимосвязях в патогенезе хронического тонзиллита, как инфекционно-аллергического процесса
0	0	52	53	Некоторые закономерности соотношения дефицита барьерной функции миндалин и системного иммунитета при хроническом тонзиллите
99	99	41	54	Прогнозування стану системи керування якістю навчального закладу
87	87	96	36	Комп’ютерна реалізація системи автоматизованого управління навчальним процесом
88	88	96	45	Формалізація вимог до системи автоматизованого управління навчальним закладом

Из таблицы можно видеть, что результаты ЛСА и LDA в некотором роде схожи, но из-за того, что LDA использует случайные величины, результаты могут быть различны на одних и тех же входных документах. Также из-за малого количества документов LDA показывает весьма большой процент схожести для документов, не подходящих к заданным ключевым словам.

Можно сделать вывод, что для проекта по извлечению публикаций из наукометрических баз данных, латентно-семантический анализ подходит лучше, нежели вероятностная модель.

ISSN 2076-2429 (print)
ISSN 2223-3814 (on line)

Из-за небольшого объема как публикаций, так и их содержимого (название в нашем случае), вероятностная модель латентного размещения Дирихле показывает худшие результаты. Учитывая, что одним из недостатков ЛСА является снижение скорости вычисления при увеличении объёма данных, для этого проекта им можно пренебречь.

Выводы. Латентное размещение Дирихле является базовой вероятностной тематической моделью и из-за большого количества приложений и обобщений является самой распространенной вероятностной тематической моделью. Базовые вероятностные тематические модели позволяют выявлять скрытую тематику документов на основе модели документа как мешка слов. В них также предполагается существование скрытых взаимосвязей между различными объектами, которые могут проявляться в структуре словоупотребления. Семантическая близость различных объектов может оцениваться путём сравнения их тематических векторов.

При применении латентного размещения Дирихле к проекту по извлечению публикаций из наукометрических баз данных замечено, что использование латентно-семантического анализа дает лучшие результаты. Поэтому, не смотря на недостатки ЛСА, использование его в этом проекте оправдано.

Литература

1. Коляда, А.С. Автоматизация извлечения информации из наукометрических баз даннях / А.С. Коляда, В.Д. Гогунский // Управління розвитком складних систем. — 2013. — Вип. 16. — С. 96 — 99.
2. Коляда, А.С. Латентно семантический подход для анализа информации из наукометрических баз даннях / А.С. Коляда // Управління розвитком складних систем. — 2014. — Вип. 17. — С. 101 —108.
3. Воронцов, К.В. Вероятностное тематическое моделирование [Электронный ресурс] / К.В. Воронцов // MachineLearning.ru. — Режим доступа: http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf (Дата обращения: 03.03.2014).
4. Daud, A. Knowledge discovery through directed probabilistic topic models: a survey / A. Daud, J. Li, L. Zhou, F. Muhammad // Frontiers of Computer Science in China. — 2010. — Vol. 4, Iss. 2. — PP. 280 — 301.
5. Blei, D.M. Latent Dirichlet Allocation / D.M. Blei, A.Y. Ng, M.I. Jordan // Journal of Machine Learning Research. — 2003. —Vol. 3. — PP. 993 — 1022.

References

1. Kolyada, A.S. Avtomatizatsiya izvlecheniya informatsii iz naukometricheskikh baz dannykh [Automating the extraction of information from scientometric databases] / A.S. Kolyada, V.D. Gogunskiy // Upravlinnia rozvytkom skladnykh system [Managing the Development of Complex Systems]. — 2013. — Iss. 16. — pp. 96 — 99.
2. Kolyada, A.S. Latentno semanticheskiy podkhod dlya analiza informatsii iz naukometricheskikh baz dannykh [Latent semantic approach for the analysis of information from scientometric databases] / A.S. Kolyada // Upravlinnia rozvytkom skladnykh system [Managing the Development of Complex Systems]. — 2014. — Iss. 17. — pp. 101 — 108.
3. Vorontsov, K.V. Veroyatnostnoe tematicheskoe modelirovanie [Probabilistic case modeling] [Electronic resource] / K.V. Vorontsov // MachineLearning.ru. Available at: http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf (Access date: 03.03.2014)
4. Daud, A. Knowledge discovery through directed probabilistic topic models: a survey / A. Daud, J. Li, L. Zhou, F. Muhammad // Frontiers of Computer Science in China. — 2010. — Vol. 4, Iss. 2. — pp. 280 — 301.
5. Blei, D.M. Latent Dirichlet Allocation / D.M. Blei, A.Y. Ng, M.I. Jordan // Journal of Machine Learning Research. — 2003. —Vol. 3. — pp. 993 — 1022.

АНОТАЦІЯ / АННОТАЦИЯ / ABSTRACT

А.С. Коляда, В.О. Яковенко, В.Д. Гогунський. **Застосування латентного розміщення Діріхле для аналізу публікацій з наукометричних баз даних**. Метою роботи є визначення найбільш відповідної тематичної моделі для класифікації наукових публікацій за автором-однофамільцем. Проаналізовано ймовірнісні тематичні моделі та запропоновано використання моделі латентного розміщення Діріхле — лідируючої серед імовірнісних тематичних

моделей завдяки численним узагальненням і додаткам до аналізу колекцій текстових документів. Для порівняння обрано модель латентно-семантичного аналізу, недоліки якої вирішуються за допомогою розглянутої моделі. Модель використана у проекті по вилученню публікацій з наукометричних баз даних. У цьому проекті застосування тематичного моделювання дозволяє вирішити проблему поділу публікацій авторів-однофамільців, де колекцією документів обрано назви публікацій. Результати показують, що модель латентного розміщення Діріхле поступається латентно-семантичному аналізу, коли використовується малий обсяг вмісту документів. Тому для колекцій документів малого обсягу переважним є використання латентно-семантичного аналізу, а для великих обсягів — латентного розміщення Діріхле.

Ключові слова: модель, латентний, семантичний, Діріхле, тематичний, публікація.

А.С. Коляда, В.А. Яковенко, В.Д. Гогунский. **Применение латентного размещения Дирихле для анализа публикаций из наукометрических баз данных.** Целью работы является определение наиболее подходящей тематической модели для классификации научных публикаций по авторам-однофамильцам. Проанализированы вероятностные тематические модели и предложено использование модели латентного размещения Дирихле — лидирующей среди вероятностных тематических моделей благодаря многочисленным обобщениям и приложениям к анализу коллекций текстовых документов. Для сравнения выбрана модель латентно семантического анализа, недостатки которой решаются при помощи рассматриваемой модели. Модель применена в проекте по извлечению публикаций из наукометрических баз данных. В этом проекте применение тематического моделирования позволяет решить проблему разделения публикаций авторов-однофамильцев, где в качестве коллекции документов выбраны названия публикаций. Результаты показали что модель латентного размещения Дирихле уступает латентно-семантическому анализу, когда используется малый обьем содержимого документов. Поэтому для коллекций документов малого обьема предпочтительным является использование латентно-семантического анализа, а для больших обьемов — латентного размещения Дирихле.

Ключевые слова: модель, латентный, семантический, Дирихле, тематический, публикация.

A.S. Kolyada, B.A. Yakovenko, V.D. Gogunsky. **Applying latent Dirichlet allocation for analysis of publications in scientometric databases.** The aim of the work is to determine the most appropriate model for a thematic classification of scientific publications by author with the same sirname. The probabilistic models are analyzed and it is proposed to use the model of latent Dirichlet allocation — the leading one among probabilistic models thanks to numerous generalizations and applications to the analysis of collections of text documents. For comparison the latent semantic analysis model is chosen. The model is used in the project for the extraction of publications from scientometric databases. In this project the usage of topic modeling solves the problem of separation of publications of authors with the same sirname, where titles of publications are selected as collection of documents. The results show that the model of latent Dirichlet allocation yield to the latent semantic analysis with usage of small volume of the contents of documents. Therefore, for small collections of documents of volume it is preferable to use latent semantic analysis, and for large volumes — latent Dirichlet allocation.

Keywords: model, latent, semantic, Dirichlet, topic, publication.

Рецензент д-р. техн. наук, проф. Одес. нац. политехн. ун-та Становский А.Л.

Поступила в редакцию 15 апреля 2014 г.