OSTIS-2015
(Open Semantic Technologies for Intelligent Systems)

УДК 004;621.398;681.5

ОБ ОДНОМ ПОДХОДЕ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Чан Ван Ан

Вьетнамский Государственный Технический Университет имени Ле Куй Дона, г. Ханой, Вьетнам

tavistu@gmail.com

В статье предлагается новый подход тематического моделирования текстов на естественном. Идея заключается в том, что, во-первых, создание метода представления текстов, где учитываются смысловые взаимосвязи между терминами, во-вторых, создание метода распределения терминов на соответствующие тематики. Смысловые взаимосвязи между терминами изображаются в виде графа, где, автором предлагается формула для определения рёбер между вершинами графа.
Ключевые слова: тематическое моделирование текстов; метод представления текстов; граф текста; латентно-семантический анализ.

Введение

В 1958 году Герхард Лисовски и Деонард Рост предложили работу по составлению каталога религиозных текстов на иврите, призванных помочь ученым определить значения терминов, которые были давно утрачены. Путём кропотливой ручной работы они собрали воедино все возможные контексты, в которых появлялся каждый из терминов. Следующей задачей было научиться игнорировать несущественные различия в формах слов и выделять те различия, которые влияют на семантику. Трудности, с которыми столкнулись Лисовски и Рост полвека назад, часто возникают и сегодня при автоматическом анализе текстов. Теоретически обоснованным и активно развивающимся направлением в анализе текстов на естественном языке, призванным решать перечисленные задачи, является тематическое моделирование коллекций текстовых документов [Коршунов, 2012].

Построение тематической модели может рассматриваться как задача одновременной кластеризации документов и слов по одному и тому же множеству кластеров, называемых темами. В терминах кластерного анализа тема — это результат би-кластеризации, то есть одновременной кластеризации и слов, и документов по их семантической близости. Обычно выполняется нечёткая кластеризация, то есть документ может принадлежать нескольким темам в различной степени. Таким образом, сжатое семантическое описание слова или документа представляет собой вероятностное распределение на множестве тем. Процесс нахождения этих распределений и называется тематическим моделированием

Тематическое моделирование текстов играет важную роль в различных областях информационной технологии, как в организации баз знаний, поиске, реферировании коллекции документов и новостных потоков, фильтрации спама и т.д. При поиске, с помощью классификации текстовых документов, можно найти наиболее эффективные информации, сократить трудозатраты на поиск нужной информации.

1. Подходы тематических моделирований

1.1. Латентно-семантический анализ (LSA)

Существуют некоторые подходы тематических моделирований. Одним из них является латентно-семантический анализ (LSA) [Landauer, 1998] для выявления структуры семантических взаимосвязей между используемыми словами.

LSA позволяет выявлять значения слов с учетом контекста их использования путем обработки большого набора текстов. Принцип действия метода заключается в том, что сравнение множества всех контекстов, в которых слова или группы слов употребляются, и контекстов, в которых они не употребляется, позволяет сделать вывод о степени близости смысла этих слов или групп слов.

В качестве исходной информации LSA использует матрицу термы-на-документы.

Элементы этой матрицы содержат веса терминов в документах, назначенные с помощью выбранной весовой функции. В качестве примера можно рассмотреть самый простой вариант такой матрицы, в которой вес термина равен 1, если он встретился в документе (независимо от количества появлений), и 0 если не встретился (Рисунок 1).

	d_1	d_2	...	...	d_{n-1}	d_n
w_1	1	1	...	...	0	1
w_2	0	1	...	...	1	1
...	...	..	..	..	..	...
...	...	...	...	...	...	...
w_m	1	0	...	...	0	1
w_{m-1}	1	1	...	...	1	0

Рисунок 1 –Матрица “термы-на-документы”.

Где d_i – документы в коллекции;

w_i – термы, извлеченные из коллекции документов.

При применении данного метода, существуют следующие недостатки:

- матрица термы-на-документы обладает большим размером, из-за того, что, коллекция документов содержит много терминов;

- не учитываются смысловые взаимосвязи между терминами.

1.2. Метод извлечения терминов из текстов

В работе [Усталов, 2012] предложен метод извлечения терминов из текстов на русском языке при помощи графовых моделей: термины являются вершинами графа, связи между ними образуются путем последовательно сканирования текста заданным окном из $N \in [2,10]$ слов. На каждой итерации для пары слов вычисляется величина связи $WC(w_1, w_2)$, обратно зависящая от расстояния между словами:

$$WC(w_1, w_2) = \begin{cases} 1 - \dfrac{d(w_1, w_2) - 1}{N - 1}, & если \quad d(w_1, w_2) \in (0, N) \\ 0 & если \quad d(w_1, w_2) \geq N, \end{cases}$$

где w_1 и w_2 - слова, $d(w_1, w_2)$ - расстояние между словами, N - размер окна.

Слова, для которых величина $WC(w_1, w_2)$ приняла нулевое значение, не включаются во множество вершин графа.

Основанием для вычисления величины $WC(w_1, w_2)$ служит наблюдения, что между двумя рядом стоящими словами часто существует семантическое отношение. Это необходимо для обеспечения связности представления текста в виде графа. Чем выше расстояние $d(w_1, w_2)$, тем ниже вероятность существования такого отношения.

При обработке графа, работа [Усталов, 2012] ведётся исключительно с одиночными именами существительными и приглагательными. Объединение этих слов в словосочетания будет выполнено на этапе сборки словосочетаний.

Данный метод обладает большим недостатком:

- учёт смысловых взаимосвязей между терминами только для тех, что находятся только в одном окне;

- не учитывается частота встречаемости пары терминов, которые находятся в одном предложении, хотя такие термины имеют смысловую взаимосвязь.

2. Новый подход к тематическому моделированию с предлагаемым методом представления текстов

Две выше работы являются наиболее близкими к данной работе. Идея заключается в том, что, во-первых, создание метода представления текстов, где учитываются смысловые взаимосвязи между терминами, и, во-вторых, создание метода распределения терминов на соответствующие тематики.

Пусть D является входной коллекцией документов (множество текстовых документов). После нормализации коллекции документов получается множество нормальных предложений, и формируется $W = \{w_i \mid i = 1...n\}$ – множество в них терминов. Каждый документ $d \in D$ представляет собой последовательность терминов $(w_{1d}, w_{2d}, ..., w_{md})$ из множества $W, (m \leq n)$. Термин может повторяться в документе несколько раз.

По мнению Воронцова К. В. [Воронцов, 2012], тема - это набор терминов, неслучайно часто совместно встречающихся в относительно узком подмножестве документов. Темы описывают содержание коллекции документов. Пусть $T = \{t_i \mid i = 1...k\}, k < n$ – множество тем в коллекции документов D. Каждая тема создается путем группирования некоторых терминов из множества терминов C - $t_i = \{w_j \mid j = 1...h\}$, $h < n$. Причем, каждый термин находится в одной теме. Это значит:

$$\begin{cases} \bigcup_{i=1}^{|T|} t_i = W \\ \bigcap_{i=1}^{|T|} t_i = \phi \end{cases}.$$

Тогда коллекция документов рассматривается в виде $D = < W, T >$.

При классификации документов, нужно решить следующие задачи:

- извлечение терминов из коллекции документов,
- группирование терминов по тематикам.

Такой процесс описывается следующей схемой (рисунок 2):

Рисунок 2 – Схема процесса группирования терминов по тематикам.

Как видно из этой схемы, в начале, осуществляется нормализация документов с целью получения слов в виде инфинитива. После чего предлагается сделать шаг извлечения терминов из нормальных документов, вместе с тем, выполняется разбиение предложений, это важный шаг для вычисления величины связей между терминами. Особенностью предлагаемой схемы является представление коллекции документов в виде графа, где, вершины графа является терминами, между ними существуют связи, которые являются ребрами графа. Как видно из схемы, после шага построения графа связей между терминами, выполняется процедура группирования терминов по тематикам.

В работе [Гринева, 2009] предложен метод извлечения ключевых терминов. Такой метод состоит из следующих шагов:

- извлечение терминов-кандидатов;
- объединение всех синонимов.

При извлечении терминов-кандидатов исходная коллекция документов разбивается на лексемы. С помощью словаря объединяются все термины, близкие по смыслу и значению.

По списку полученных терминов на предыдущем шаге, строится семантический граф, вершинами которого являются термины документа, наличие ребра между двумя вершинами означает тот факт, что термины семантически связаны между собой, вес ребра является численным значением семантической близости двух терминов, которые соединяет данное ребро. Граф предлагается в виде $G = < W, E >$, где W - множество терминов коллекции документов D, а E - связи между ними. Предлагается, что чем ближе по смыслу, тем меньше по расстоянию.

Для определения связей между терминами, предлагаются следующие гипотезы:

- учет взаимосвязи между терминами;
- термины, которые находятся в одном предложении, имеют связь по смыслу;

Для оценки величины связи предлагается параметр $P(w_i, w_j)$, значение данной оценки тем меньше, чем больше смысловая взаимосвязь между терминами w_i и w_j ;

частота встречаемости пары терминов w_i и w_j тем больше, чем больше смысловая взаимосвязь между терминами w_i и w_j .

Из этих гипотез автором предлагается формула для определения рёбер между вершинами графа:

$$P(w_i, w_j) = \begin{cases} \prod_{k=1}^{F} \frac{\min[s_k(w_i, w_j)]}{L_k}, \text{если } w_i \text{ и } w_j \\ \text{в одном предложении} \\ 0, \text{ если } w_i \text{ и } w_j \text{ ни раз} \\ \text{не находятся в одном предложении} \end{cases} \quad (1)$$

где:

F - частота встречаемости пары терминов w_i и w_j в предложениях;

$s_k(w_i, w_j)$ - позиционная мера пары терминов w_i и w_j в k-ом предложении;

L_k - длина k-ого предложения.

Как отмечено в работе [Усталов, 2012],

$s_k(w_i, w_j) = \left| p_k(w_j) - p_k(w_i) \right|$, при w_i # w_j , где $p_k(w_i)$ - порядковый номер термина w_i в k-ом предложении. Чем выше расстояние

$s_k(w_i, w_j)$, тем ниже вероятность существования такого отношения.

Если в k-ом предложении существуют несколько пар терминов c_i и c_j, то выбирается пара, которая имеет самую маленькое значение позиционной меры - $\min[s_k(w_i, w_j)]$.

3. Анадиз результата

По данному методу проведем анализ следующих текст:

Перый текст:

В статье предлагаются модели и средства для интерактивного взаимодействия в процессе обучения неродному языку.

Описывается четыре вида интерактивного взаимодействия: интерактивное взаимодействие между членами курса при обучении теоретических основ курсов, интерактивность при взаимодействии между пользователями по принципу «социальной сети», интерактивность при проведении практики, интерактивное взаимодействие между учащимися и средствами системы при самообучении.

Второй текст:

В статье предлагается математическая модель процесса обучения неродному языку, как процесса происходящего с одновременным присутствием двух противоположностей: наличием родного языка и неродного языка. Системы с такими свойствами называются амбивалентными.

Для реализации обучающей среды предлагается технологическая схема обучения.

Трьетий текст:

В статье предлагается математическая модель амбивалентной системы обучения неродному языку, в которой отражаются такие особенности процесса обучения как появление отката и окостенения.

На основе применения численных методов таких, как метод золотого сечения и метод Тьюки определяются моменты времени начала этих явлений. Предложенное в данной статье решение этой задачи позволит увеличить эффективность процесса обучения неродному языку.

После нормализации, (в данной работе рассматриваются только существительные и прилагательные), получается следующий нормальный фрагмент с семантическим графом:

Перый текст:

Статья| математический| модель| процесс| обучение| неродной| язык|, процесс| противоположность| родной| язык| неродной| язык|. Система| свойство| амбивалентный|. Реализация| обучающий| среда| технологический| схема| обучение|.

Второй текст:

Статья| математический | модель| процесс| обучение| неродной| язык|, процесс| противоположность| родной| язык| неродной| язык|. Система| свойство| амбивалентный|. Математический| модель| дифференциальный| уравнение| Колмогоров|. Модель| обучение| основа| уравнение| зависимость| уровень| знание| обучаемый| начало| уровень| знание|, интенсивность| процесс| изучение| неродной|

Трьетий текст:

Статья| математический| модель| амбивалентный| система| обучение| неродной| язык| особенность| процесс| обучение| появление| откат| окостенение.| Основа| применение| численный| метод| метод| золотой| сечение| метод| Тьюки| момент| время| начало| явление.| Предложенный| данный| статья| решение| задача| эффективность| процесс| обучение| неродной| язык.

Используя формулу (1) получаем следующий семантический граф терминов:

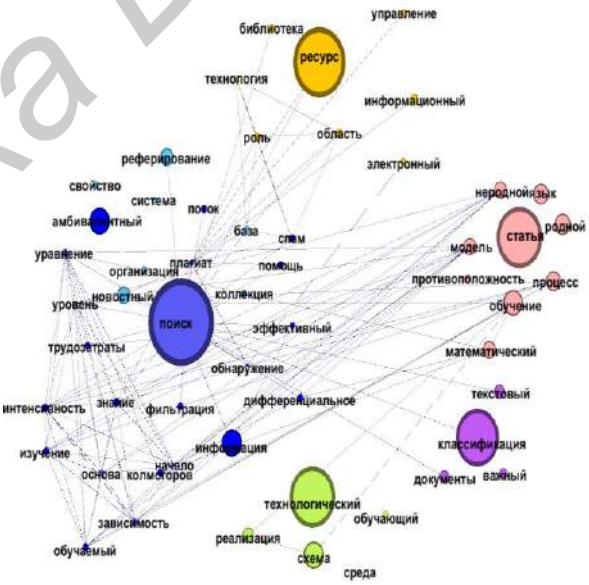

Рисунок 3 - Семантический граф терминов.

После получения семантического графа терминов выполняется шаг автоматического обнаружения кластеров терминов. При решении данной задачи был выбран метод K-medoid [Гринева, 2009] для распределения набора объектов графа (терминов) по k группам, таким образом, чтобы конечная сумма несхожести между каждым объектом была минимальна. То есть, используется критерий абсолютной погрешности. По такому методу, для выше приведенного примера, если данная коллекция документов распределяется по 4 тематикам, то получается следующие таблицы 1 и 2 распределения терминов.

Таблица 1. Распределения терминов на 4 тематик

Тематика 1	Тематика 2
неродной уровень знание интенсивность процесс язык обучение система амбивалентный свойство эффективность особенность появление откат окостенение противоположность родной дифференциальный уравнение зависимость обучаемый изучение колмогоров	момент начало явление время метод золотой сечение тьюки численный применение основа

Таблица 2. Распределения терминов на 4 тематик

Тематика 3	Тематика 4
технологический реализация обучающий среда схема	статья модель математический решение задача

При распределении на 10 тематик получается таблица 3, 4, 5, 6 и 7:

Таблица 3. Распределения терминов на 10 тематик

Тематика 1	Тематика 2
процесс модель обучение появление откат окостенение особенность противоположность изучение	момент время
дифференциальный колмогоров эффективность задача	

Таблица 4. Распределения терминов на 10 тематик

Тематика 3	Тематика 4
технологический реализация обучающий среда схема	статья решение математический

Таблица 5. Распределения терминов на 10 тематик

Тематика 5	Тематика 6
тьюки золотой метод сечение	начало уровень зависимость уравнение обучаемый явление

Таблица 6. Распределения терминов на 10 тематик

Тематика 7	Тематика 8
применение численный основа	амбивалентный свойство система

Таблица 7. Распределения терминов на 10 тематик

Тематика 9	Тематика 10
интенсивность знание	неродной родной язык

Как видно из полученных таблиц, термины в одной тематике имеют сильные смысловые взаимосвязи, такие термины характеризуют содержательный смысл тематик, в которых они находятся.

Заключение

В статье был представлен новый подход тематического моделирования текстов на естественном, которое является перспективным инструментом для обработки больших коллекций документов.

Данный подход позволяет автоматически систематизировать и реферировать электронные архивы большого масштаба, который человек не в

силах обработать, даёт возможность использовать ресурсы Интернет нетолько как корпус текстов, но и как полноценную базу данных.

Библиографический список

[Коршунов, 2012] Коршунов, А. В. Тематическое моделирование текстов на естественном языке / А. В. Коршунов, А. Г. Гомзин // Труды Института системного программирования РАН, 2012, Том 23, С. 215-242.

[Landauer, 1998] Landauer, T. An introduction to latent semantic analysys / T. Landauer, P. Foltz, and D. Laham // In Discourse Processes, 1998, v. 25, pp. 259-284.

[Усталов, 2012] Усталов, Д. А. Извлечение терминов из русскоязычных текстов при помощи графовых моделей/Д. А. Усталова// Теория графов и приложения. Материалы конференции. С. 62-69.

[Воронцов, 2012] Воронцов, К.В. Регуляризация, робастность и разреженность вероятностных тематических моделей / К.В. Воронцов, А.А. Потапенко// Компьютерные исследования и моделирование, 2012, Том 4, № 4.

[Гринева, 2009] Гринева, М. Анализ текстовых документов для извлечения тематически сгруппированных ключевых терминов / М. Гринева, М. Гринев // Труды Института системного программирования РАН, 2009, Том 16, С. 155-165.

ABOUT AN APPROACH OF TOPIC MODELING FOR TEXT IN NATURAL LANGUAGE

Tran Van An *

** Le Quy Don University of Science and Technology, Ha Noi, VietNam*

tavistu@gmail.com

Introduction

The paper proposes a new approach of topic modeling for text in natural language. Firstly, the creation a method for submission of texts. Secondly, the creation of the distribution terms for each relevant topics. Semantic relationships between terms are represented as a graph. The author proposes a formula for define the edges between the vertices of the graph.

Main part

Constructed semantic graph whose vertices are the terms of the document. Terms semantically related to each other, the weight of an edge is the numerical value of the semantic proximity of the two terms, which connects the given edge.

After receiving the semantic graph of terms is performed step automatic detection of clusters of terms. In solving this problem was chosen method of K-medoid for the distribution of a vertex set of the graph (terms) in groups.

Conclusion

The paper was presented new approach of topic modeling for text in natural language, which is a promising tool for processing large document collections. This approach supports you to automatically organize and abstracted electronic archives of a large scale, which people can not handle, makes it possible to use Internet resources not only as corpus, but also as a full-fledged database.