Анализ коллекции нормативных документов 2007 года средствами системы SOPHIA

© В.Ю. Добрынин

СПбГУ, Sophia Search
v.dobrynin@bk.ru

Аннотация

Коллекция нормативных документов 2007 года кластеризована в 659 кластеров методом контекстной кластеризации. Утверждается, что полученные кластеры представляют дискурсивные сообщества, представленные в данной коллекции документов. Представлены иллюстративные материалы в виде кратких описаний некоторых кластеров различных размеров, а также статистические данные по всему набору кластеров.

1. Введение

Метод контекстной кластеризации (Contextual Document Clustering, CDC) [4] выделяет кластеры текстовых документов, которые можно интерпретировать как совокупности текстов, порожденных некоторыми дискурсивными сообществами (например различными профессиональными группами). Различные дискурсы могут отличаться функциональными стилями, что характеризуется лексическими особенностями (жаргонизмы, терминология) [1].

Метод CDC основан на анализе коллекции текстовых документов, в процессе которого происходит выявление жаргонизмов и терминологии специфичной для различных дискурсивных сообществ. Далее контексты выделенных терминов играют роль аттракторов кластеров, собирая вокруг себя тексты, порожденные определенным сообществом.

Метод контекстной кластеризации был протестирован на стандартных тестовых коллекциях (Reuters-21578, Reuters-RCV1, OHSUMED) [5-11] и применялся при кластеризации больших коллекций документов в реальных приложениях: рефераты патентов – 4,500,000 документов, и рефераты статей (Medline) – 18,000,000 документов.

В данной работе описываются результаты применения CDC к коллекции нормативных документов 2007 года [2].

2. Метод контекстной кластеризации

В основе данного метода лежит очевидная идея о том, что авторы текстов, представленных в данной коллекции документов, могут быть разбиты на группы (возможно пересекающиеся), относящиеся к различным профессиональным, социальным, культурным и т.п. сообществам (дискурсивное сообщество). В каждом сообществе используется свой особый язык, что выражается, например, в использовании терминов специфических для данного сообщества.

Для поиска таких терминов выполняется построение контекстов для всех слов словаря коллекции за исключением очнь редких и очень популярных слов. Под контекстом слова понимается распределение вероятностей всех слов, которые встречаются совместно с данным словом в одном документе. Иными словами, контекст слова z определяется как

$$p(y \mid z) = \frac{\sum_{x \in \mathrm{X}(z)} tf(x,y)}{\sum_{x \in \mathrm{X}(z), t \in \mathrm{Y}} tf(x,t)},$$

где $p(y \mid z)$ есть вероятность слова y в контексте слова z, $tf(x,y)$ есть частота встречаемости слова y в документе x, $\mathrm{X}(z)$ есть множество всех документов, содержащих слово z, и Y есть словарь коллекции.

Слова, специфические для отдельных сообществ, должны иметь контексты с относительно невысокой энтропией. Выбор таких слов основан на вычислении энтропии всех построенных контекстов с учетом частоты встречаемости слов в коллекции. Учет частоты встречаемости слова связан с тем, что сообщества могут быть представлены в данной коллекции как большим, так и малым числом документов, и учет одной только энтропии приведет к преимущественному выделению слов, специфических для малых сообществ.

Контексты, соответствующие выделенным специфическим словам, могут рассматриваться как темы, представленные в коллекции. Однако представляется, что использование термина «тема» не является удачным в данном случае. Используя полученные контексты в качестве аттракторов кластеров, можно выполнить кластеризацию коллекции, связывая каждый документ с ближайшим к нему контекстом (в данных экспериментах использовалась дивергенция Jensen-Shannon для оценивания расстояния между контекстом и вероятностным распределением слов в документе). Множество документов одного кластера можно рассматривать как дискурс соотвествующего сообщества авторов. Характерной особенностью кластера данного типа должно быть то, что наиболее важные для соотвествующего дискурса слова должны встречаться в документах из этого кластера в одном и том же значении. Это следует из соображений экономии усилий – члены одного достаточно узкого сообщества не должны тратить время на выяснение смысла используемых слов, которые вполне могут быть многозначными во «внешнем» мире, но должны восприниматься однозначно в данном сообществе.

Наряду с собственно алгоритмом кластеризации CDC, на котором и базируется система SOPHIA [3], в рамках данной системы реализован ряд алгоритмов выявления значимых слов и фраз, которые могут использоваться при построении описаний как отдельных документов, так и их групп (релевантных запросу) и кластера в целом.

3. Описание построенной системы кластеров

3.1 Основные характеристики построенных кластеров

В процессе индексирования текста не применялся стемминг или какой-либо морфологический анализ и не удалялись стоп-слова. Не применялись никакие внешние словари и иные ресурсы. В следующей таблицы приведены основные хорактеристики построенной системы кластеров.

Число документов	319,555
Размер словаря коллекции	678,211
Число кластеров	659
Максимальный размер кластера	17,122
Средний размер кластера	484

Медиана размера кластера	220
Максимальный размер словаря кластера	85,325
Средний размера словаря кластера	9,904
Медиана размера словаря кластера	6,780

3.2 Иллюстративный материал

Имеются различные вырианты построения описаний кластеров, как статические, так и динамические, зависящие от множества релевантных документов, находящихся в данном кластере. В данной работе в качестве статического описания содержимого кластера предлагается набор значимых фраз, извлеченных из всей совокупности текстов данного кластера. В последующих таблицах представлены примеры описаний кластеров большого, среднего и малого размеров.

Кластер большого размера

кластер	Число документов	Фраза
37403	10,046	Кремль
		указ президента российской федерации о награждении орденом дружбы
		указ президента российской федерации о награждении орденом за заслуги перед отечеством
		президент российской федерации б
		президент российской федерации в
		официальный электронный текст нтц система
		ельцин москва
		наградить орденом
		путин москва
		Большой вклад в

Кластер среднего размера

кластер	Число документов	фраза
126833	941	распределительными газопроводами и дифференцированные по группам потребителей тарифные
		услуги по транспортировке газа по газораспределительным сетям
		услуги по транспортировке газа по распределительным газопроводам
		куб
		по поставке транспортировке газа по распределительным газопроводам
		по поставке транспортировке газа по газораспределительным сетям
		тарифная ставка
		вводимые в действие с
		фст россии от
		года постановление фэк россии от

Кластер малого размера

кластер	Число документов	фраза
33083	100	внутренних войск мвд россии
		гражданского персонала железнодорожных войск
		деятельности железнодорожных войск
		железнодорожных войск и
		федеральной службы железнодорожных войск российской федерации командующего
		командующий железнодорожными войсками
		федерации командующего железнодорожными войсками российской

		федерации
		войск гражданской обороны российской
		обороны российской федерации
		о железнодорожных войсках российской федерации
		в войсках
		министерства обороны

Приведенные примеры описаний кластеров показывают наличие определенных технических проблем, связанных с определением границ предложений. Однако выбор ключевых слов, вокруг которых и формируются предложения, представляется вполне удачным.

В следующей таблице приведены заголовки документов, находящихся как вблизи аттрактара, так и на максимальном удалении от аттрактора для каждого из ранее упомянутых кластеров. Эти примеры демонстрируют тематическую однородность полученных кластеров.

кластер	документ	расстояние	заголовок
37403	299974	0.23684	О назначении Юхина В.В. Чрезвычайным и Полномочным Послом Российской Федерации
37403	297388	0.75578	О награждении государственными наградами Российской Федерации военнослужащих Пограничных войск Российской Федерации
126833	80544	0.14414	О тарифах на услуги по транспортировке газа по распределительным газопроводам Омской области (с изменениями на 25 декабря 2003 года) (утратило силу с 01.09.2005 на основании приказа ФСТ России от 23.08.2005 N 391-э/4)
126833	73274	0.68638	О рассмотрении ходатайств открытых акционерных обществ "Ейскгоргаз", "Абинскрайгаз", "Армавиргоргаз",

			"Белаяглинарайгаз", "Белореченскрайгаз", "Брюховецкаярайгаз", "Геленджикрайгаз", "Динскаярайгаз", "Калининскаярайгаз", "Красноармейскаярайгаз"...
33083	244931	0.2106	Вопросы Железнодорожных войск Российской Федерации (с изменениями на 3 апреля 2002 года) (утратил силу на основании Указа Президента РФ от 03.08.2005 N 918)
33083	271691	0.67758	О порядке проведения профилактики арендованных каналов связи

Из заголовка последнего документа в предыдущей таблице не видно, что данный документ релевантен тематике соответствующего кластера, однако знакомство с содержимым документа позволяет утверждать, что этот документ также является релевантным (в документе рассматривается использование каналов связи внутренними войсками).

Литература

[1] А.А. Кибрик. Модус, жанр и другие параметры классификации дискурсов. *Вопросы языкознания*, в печати, 2009

[2] Коллекция нормативных документов 2007. http://romip.ru/ru/collections/legal07.html

[3] Система анализа текстовых коллекций и поиска SOPHIA. http://www.sophiasearch.com

[4] V. Dobrynin, D. W. Patterson, N. Rooney. Contextual Document Clustering. ECIR 2004, pages 167-180, 2004

[5] V. Dobrynin, S. K. Pham, D. Patterson, N. Rooney, M. Galushka: SOPHIA in Enterprise Track. TREC 2006

[6] V. Dobrynin, D. W. Patterson, M. Galushka, N. Rooney. SOPHIA: an interactive cluster-based retrieval system for the OHSUMED collection. In *IEEE Transactions on Information Technology in Biomedicine*, volume 9(2), pages 256-265, 2005

[7] D. Patterson, N. Rooney, M. Galushka, V. Dobrynin, E. Smirnova. SOPHIA-TCBR: A knowledge discovery framework for textual case-

based reasoning. In *Knowl.-Based Syst*. Volume 21(5), pages 404-414, 2008
[8] D. W. Patterson, N. Rooney, V. Dobrynin, M. Galushka. Sophia: A novel approach for Textual Case-based Reasoning. IJCAI 2005, pages 15-20, 2005
[9] Niall Rooney, David W. Patterson, Mykola Galushka, Vladimir Dobrynin, Elena Smirnova: An investigation into the stability of contextual document clustering. In *JASIST*, volume 59(2), pages 256-266, 2008
[10] N. Rooney, D. W. Patterson, M. Galushka, V. Dobrynin. A relevance feedback mechanism for cluster-based retrieval. In *Inf. Process. Manage*. volume 42(5), pages 1176-1184, 2006
[11] N. Rooney, D. W. Patterson, M. Galushka, V. Dobrynin. A scaleable document clustering approach for large document corpora. In *Inf. Process. Manage*., Volume 42(5), pages 1163-1175, 2006

SOPHIA: Analysis of a Legal Document Collection

Vladimir Dobrynin

The legal document collection was clustered into 659 clusters by Contextual Document Clustering algorithm. Interpretation of generated cluster structure based on concept of discourse community is presented.