УДК 519.21:681.142

АВТОМАТИЗИРОВАННАЯ КЛАССИФИКАЦИЯ НА ОСНОВЕ ЛАТЕНТНО-СЕМАНТИЧЕСКОГО АНАЛИЗА

Помилуйко П.А.,
аспирант МНУЦ ИТС

pomiluyko@gmail.com

Статья посвящена описанию подходов, используемых для классификации коллекции текстовых документов.
Ключевые слова и фразы: автоматизация классификации, поисковая система, извлечение информации, анализ текста, латентно-семантический анализ.

This article describes approaches are used for classification of document collection.
Key words: automated classification, search engine, information retrieval, text analysis, latent semantic analysis.

Стаття присвячена опису підходів, які використовуються для класифікації колекції текстових документів.
Ключові слова: автоматизація класифікації, пошукова система, вилучення інформації, аналіз тексту, латентно-семантичний аналіз.

1. Актуальность

Область современного информационного поиска чрезвычайно разнообразна. Она включает такие задачи, как собственно поиск информации, фильтрация, рубрикация и кластеризация документов, поиск ответов на вопросы, автоматическое аннотирование документа и группы документов, поиск похожих документов и дубликатов и многое другое. Когда подобные операции выполняет человек, ему необходимо определить основное содержание документа, его основную тему и подтемы, и для этого обычно используется большой объем знаний о языке, мире, организации связного текста.

Латентно-семантический анализ (ЛСА) — это метод обработки информации на естественном языке, анализирующий взаимосвязь между коллекцией документов и терминами в них встречающимися, сопоставляющий некоторые факторы (тематики) всем документам и термам. В основе метода латентно-семантического анализа лежат принципы факторного анализа, в частности, выявление латентных связей изучаемых явлений или объектов. С помощью факторного анализа возможно выявление скрытых переменных факторов, отвечающих за наличие линейных статистических связей корреляций между наблюдаемыми переменными.

Актуальность его использования вызвана экспоненциальным ростом объема информации и низким качеством точности поиска. Главные цели использования ЛСА – выявление семантических связей между термами и

латентных зависимостей внутри множества текстовых документов, распределения (классификации) документов на группы, расширения поисковых запросов, и некоторых других задач.

Одной из проблем широкого использования ЛСА в информационных системах является его высокая сложность и значительное снижение скорости вычисления при увеличении объема входных данных. В работе было проведено исследование моделей и методов, которые могли бы помочь оптимизировать его использование.

2. Анализ последних исследований

ЛСА позволяет выявлять значения слов с учетом контекста их использования путем обработки большого набора текстовой информации. Впервые метод ЛСА был описан в работе [4] и затем развит в трудах Scott Deerwester, Susan Dumais, George Furnas и др.

Представления слова и абзаца с помощью метода ЛСА во многом моделируют восприятие текста человеком [5]. Например, с его помощью можно оценить эссе на соответствие теме или сопоставить смыслы отрывков текста. Кроме того, содержательные примеры работы метода ЛСА могут быть найдены в работах [4], [8].

3. Задачи и цели исследований

Предположим, стоит задача написать алгоритм, который сможет отличать текстовые документы по их тематической принадлежности. Первое, что приходит в голову, это выбрать слова которые встречаются исключительно в документах каждого вида и использовать их для классификации. Очевидная проблема такого подхода: как перечислить все возможные слова и что делать в случае, когда в статье есть слова из нескольких классов. Дополнительную сложность представляют омонимы. Т.е. слова имеющие множество значений. Например, слово «банки» в одном контексте может означать стеклянные сосуды, а в другом контексте это могут быть финансовые институты. Латентно-семантический анализ отображает документы и отдельные слова в так называемое «семантический образ», в котором и производятся все дальнейшие сравнения. Не стоит также забывать о сложности вычислений выполнения таких сравнений.

Сегодня ЛСА находят лишь ограниченное применение в поисковых системах. Большинство информационных поисковых систем не используют онтологию. Задачу поиска традиционно решают на основе изощренных методов учета частоты встречаемости слов в тексте, расстоянии между словами и т. п. Хотя задачу поиска можно существенно упростить, классифицируя полученные результаты на тематические разделы. Еще один существенный недостаток заключается в том, что в основном используются только простые

слова, хотя очень часто словосочетания могут нам сказать намного больше, чем слова по отдельности. Такие методы имеют ряд очевидных недостатков, которые затрудняют поиск релевантных текстов.

В этой статье рассматриваются методы построения классификатора на основании латентно-семантического анализа. Каждый раздел классификатора описывается семантическим множеством термов.

4. Исследования классификации коллекции документов

Представляемая технология базируется на так называемом дедуктивном подходе к составлению классификатора, который подразумевает, что все дескрипторы (слова и словосочетания) извлекаются из коллекции документов. При ручном составлении, эксперт нередко добавляет термины не присутствующие в корпусе, однако количество подобных слов и словосочетаний, как правило, не превышает 20-25%. Особенности предлагаемой технологии позволяют эксперту добавить недостающие термины вручную, при необходимости.

Главное отличие от ручного составления заключается в том, что производится предварительная машинная обработка коллекции документов и эксперту на каждом из этапов конструирования классификатора представляются множество слов и словосочетаний. Согласно технологии человек всегда сам определяет окончательный список элементов, имея возможность добавления, удаления и модифицирования списка кандидатов, предложенных системой. Технология автоматизированного построения классификатора состоит из нескольких этапов:

1) Определение основных разделов классификатора. Первоначально разделы – таксоны первого уровня и основные дескрипторы задает эксперт. Это может быть разделы энциклопедии, справочника и т.д. Главным принципом составления названий является общая тематика всех разделов. Для улучшения качества результатов также рекомендуется изначально указать для каждого раздела несколько основных дескрипторов.
2) По каждому названию таксона производиться анализ через Википедию и API Google Search. Это необходимо для того, чтобы предварительно собрать как можно больше информации о разделах, которые будут использоваться как таксоны 1-го уровня в классификаторе. Этот этап в данной статье подробно рассматриваться не будет.
3) Токенизация – обработка коллекции документов. Выполняются простейшие, но необходимые преобразования коллекции документов с целью представления их в виде пригодном для дальнейшей обработки с помощью методов машинного анализа текстовой информации.

4) Построение множества предпочтительных дескрипторов. Формирование множества слов и словосочетаний кандидатов для включения в семантический образ. Для каждого дескриптора вычисляется его частота встречаемости (весовой коэффициент).
5) Поиск в словаре синонимов отношений связей между дескрипторами, редактирование экспертом автоматически найденных отношений и окончательное группирование дескрипторов. Эксперт руководствуясь множеством кандидатов составляет список ключевых понятий предметной области – семантический образ.

Большая коллекция может исчисляться в десятках, сотнях тысяч, а иногда и миллионах документов. Количество слов в таких корпусах текстов нередко достигает десятков миллионов. При этом, текст написанный на естественном языке слабо структурирован и может содержать большое количество ошибок. Выделение значимых слов и словосочетаний в такой коллекции и применение алгоритмов анализа текстов требует предварительной обработки коллекции. Для оптимизации работы были выбраны следующие этапы предварительной обработки текста:

4.1. Извлечение всех слов и словосочетаний из корпуса текстов

Множество всех уникальных слов содержащихся в корпусе T формируется с помощью процедуры токенизации. Алгоритм определят границы слов с помощью множества стоп-знаков слова – множества знаков, которые позволяют отделять в тексте слова друг от друга [7], а также некоторых правил, после чего составляется список всех уникальных слов в корпусе. Размер множества T, как правило, не превышает 50000 слов. Будем считать также, что T содержит пустую строку.

Важно извлечь из текста не только отдельные слова, но и словосочетания т.к. основные понятия предметной области очень часто представлены составными словами. К примеру, словосочетание «коммерческая недвижимость» скажем нам гораздо больше, нежели слова «коммерческая» и «недвижимость» по отдельности. По соображениям целесообразности, мы ограничиваем максимальную длину искомых дескрипторов четырьмя словами. При этом, отдельный дескриптор d можно представить как 4-компонентный упорядоченный кортеж (w1, w2, w3, w4) є T.

Для поиска множества значимых словосочетаний были применены конечные преобразователи (finite state transducers), использующие специально подготовленные словари терминов, из которых может быть составлено словосочетание [3]. В результате данного этапа формируется множество дескрипторов корпуса D = W U MWE.

4.2. Стемматизация слов

Стемматизация (стемминг) – это процесс нахождения основы слова для заданного исходного слова. Следует обратить внимание, что основа слова необязательно совпадает с морфологическим корнем слова. В этом исследовании мы использовали стеммер Портера [n]. Алгоритм не использует баз основ слов, а лишь, применяя последовательно ряд правил, отсекает окончания и суффиксы, основываясь на особенностях языка, в связи с чем работает быстро, но не всегда безошибочно. Этот этап выполняется для того чтобы связать одинаковые слова текста находящиеся в разных словоформах.

4.3 Установление частей речи

Установление частей речи извлеченных слов и выражений производится с помощью специального модуля в синтаксическом анализаторе, который назначает каждому слову соответствующую часть речи. Данный этап делает возможным фильтрацию слов и словосочетаний по части речи, так, к примеру, из слов кандидатов в дескрипторы исключаются глаголы и глагольные конструкции, а предпочтение отдается существительным.

4.4 Декапитализация, деакцентизация и удаление стоп-слов

Декапитализация – преобразование всех символов корпуса к нижнему регистру. Деакцентизация – прием используемый при обработке текстов на французском языке, в которых существуют буквы с акцентами, к примеру é, è, à и т.п. Из-за особенностей грамматики, одно и то же слово в разных контекстах появляться с акцентами и без, поэтому все символы с акцентами заменяются на аналоги без акцентов. Стоп лист – список вспомогательные слов несущих мало информации о содержании документа, таких как артикли, союзы и наиболее распространенные глаголы. Из текста удаляются все слова из стоп листа.

4.5. Вычисления частоты встречаемости дескрипторов

На данном этапе вычисляется частота встречаемости дескрипторов с учетом заголовков и названий документов. Для этого используется статистическая мера TF-IDF [n], используемая 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.

После выполнения всех этапов обработки текста для каждого раздела строится множество предпочтительных дескрипторов с учетом частоты встречаемости слов в коллекции. Эксперт утверждает дескрипторы из которых образуется семантический образ. Чем больше дескрипторов будет выбрано для каждого раздела, тем выше вероятность правильного определения раздела. В следствии, каждый документ коллекции (а точнее его образ) сравнивается с семантическим образом каждого раздела. Образ каждого документа состоит из дескрипторов, которые мы получили на этапе предварительной обработки текста. В случае, когда сумма весовых коэффициентов дескрипторов семантического пространства, которые также существуют в образе документа, превышает пороговое значение (вычисляется эмпирически) – документ коллекции относится к соответствующему разделу.

5. Заключение

Классификация коллекции документов существенно упрощает процесс поиска. Но для того, чтобы корректно распределить коллекцию документов по тематической принадлежности требуется немало усилий человеческого труда. Автоматизация этого процесса существенно поможет в решении данной задачи. На данном этапе автоматизировано построение семантических образов при условии наличии первоначальных признаков (хотя бы 4-5 дескрипторов). В дальнейшем планируется автоматизировать весь процесс построения. В дальнейшем планируется оптимизировать алгоритм образования семантического пространства. Это направление является весьма перспективным и будет приоритетным в дальнейшем.

Список использованной литературы

1. OpenCyc, http://www.opencyc.org.
2. The Dublin Core Metadata Initiative (DCMI), http://dublincore.org
3. Karttunen L. (2000). «Applications of Finite-State Transducers in Natural Language Processing». 5th International Conference on Implementation and Application of Automata.
4. Baeza-Yates R., Ribeiro-Neto B. (1999). «Modern Information Retrieval». Addison Wesley Longman Publishing Co. Inc.
5. Frakes W., Baeza-Yates R.(1992), «Information Retrieval. Data Structures & Algorithms». Prentice Hall PTR; Facsimile edition.
6. Aitchison. J.(2002) Thesaurus Construction and Use: A Practical Manual. Routledge, 4 edition.
7. American National Standards Institute. ANSI/NISO Z39.19-2005: Guidelines

for the Construction, Format, and Management of Monolingual Controlled Vocabularies, 2005.

8. Peirsman Y., Heylen K., Speelman D. (2007). «Finding semantically related words in Dutch. Co-occurrences versus syntactic contexts». In Proceedings of the CoSMO workshop, Roskilde, Denmark.
9. Филиппович Ю., Прохоров А.(2002) «Семантика информационных технологий: опыты словарно-тезаурусного описания». Изд-во МГУП, ISBN 5-8122-0367-9.
10. Berry M., Dumais S., O'Brien G. (1994). «Using Linear Algebra for Intelligent Information Retrieval». Society for Industrial and Applied Mathematics. Berry M., Drmac Z.,Jessup R. (1999).Matrices, Vector Spaces, and Information Retrieval. Society for Industrial and Applied Mathematics.
11. Sahlgren M. (2006). «The Word-Space Model. Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces». A PhD dissertation submitted to Stockholm University.
12. Gregory G. (1994). «Explorations in Automatic Thesaurus Discovery». The Springer International Series in Engineering and Computer Science.
13. Добров Б.В., Лукашевич Н.В., Сыромятников С.В. (2003). «Формирование базы терминологических словосочетаний по текстам предметной области».