ИНТЕГРИРОВАННЫЙ ПОДХОД ТЕКСТОВОЙ КЛАСТЕРИЗАЦИИ ДЛЯ НЕСТРУКТУРИРОВАННЫХ ДОКУМЕНТОВ

Татьяна Шатовская[1], Ирина Каменева[2]

[1]Харьковский Национальный Университет Радиоэлектроники
ул. Балакирева, 50,17, Харьков, 61000, Украина, тел.: (8057) 702-16-46 E-mail: tanita_uk@mail.ru
[2]Харьковский Национальный Университет Радиоэлектроники
ул. Клочковская, 193, Харьков, 61141, Украина, тел.: (8057) 702-16-46, E-mail:irina.kamenieva@gmail.com

Аннотация

В данной статье мы представляем интегрированный подход к классификации текста для структурированного хранения документов на компьютере. Данный подход основан на двух методах иерархической кластеризации, которые относятся к области text и data mining. Первым этапом является предварительная обработка документов, что позволяет сократить время и качественно вычислить результат. После предварительной обработки мы используем векторную модель, которая позволяет четко определить значимость слов в документе. Заключительным этапом является иерархическая кластеризация, в которую входят два метода dendrogramms и k-means. Метод дендрограммы позволяет предварительно определить количество кластеров (папок), метод к-средних относит документы к определенным кластерам. Завершающим этапом так же является применение метода дендрограммы для создания иерархической последовательности документов внутри каждого кластера (папки).

Вступление

В настоящее время классификация текста является одной из актуальных научно – исследовательских проблем. Методы классификации текстов применяются в фильтрации документов, распознавании спама, автоматическом аннотировании, снятии неоднозначности (автоматические переводчики), составлении Интернет каталогов, классификации новостей, распределении рекламы, персональных новостях.

Классификация документов применяется во многих приложениях: фильтрация электронной почты, маршрутизация процесса пересылки почты, фильтрация спама, контролирование новостей, избирательное распространение информации для потребителей, автоматическая индексация научных статей, автоматическая популяция иерархических категорий Web – ресурсов, идентификация документационного жанра, компетенция авторства, исследование кодирования.

С каждым годом увеличивается объем доступных пользователю массивов текстовой информации на рабочем компьютере, что способствует веще большей актуализации задачи поиска необходимых пользователю документов в таких массивах. Для решения этой задачи часто применяются различные тематические классификаторы, рубрикаторы и т. д., которые позволяют искать (автоматически или вручную) документы в небольшом подмножестве документной базы, соответствующем интересующей пользователя тематике[1].

В данной статье рассматривается метод автоматической кластеризации, который способен без помощи пользователя раскластеризировать хаотическое расположение файлов в структурированный набор документов относительно тематики [2].Такая система поможет хранить информацию на компьютере в надлежащем виде и не затрачивать время на разбор документов по соответствующим тематикам.

Постановка задачи

На сегодняшний день существует очень большое количество классификаторов. Чаще всего используются методы Байесовской классификации, метод опорных векторов и многие другие. Эти методы имеют существенный недостаток – обучение с учителем (supervised learning). В данном случае классификация определяет документы в одну или более предопределённых категорий. В классификации текста новый текстовый документ назначается в один из уже существующих наборов документов класса. В нашей статье мы используем текстовую кластеризацию – обучение без учителя (unsupervised learning). Она применяется для нахождения новых заранее не известных структуры. [3] Текстовая кластеризация автоматически выявляет группы семантически похожих документов среди заданного фиксированного множества документов. Следует отметить, что группы формируются только на основе попарной схожести описаний документов, и никакие характеристики этих групп не задаются заранее, в отличие от текстовой классификации, где категории задаются заранее.[4]

На каждом рабочем компьютере существует огромное количество папок, в которых зачастую хранится большое количество документов, которые, как правило, имеют абсолютно разную тематику. Человек по истечении определенного промежутка времени с трудом вспоминает, что в какой папке

находится, а если промежуток перетекает в месяцы, то вообще невозможно вспомнить, в какой папке храниться необходимая ему на данный период информация годичной давности. Наша система позволяет решить данную проблему. Она автоматически кластеризирует документы в папки, которые соответствуют тематике документа. Пользователю необходимо будет воспользоваться предложенной системой и документы отнесутся к логическим по структуре документа папкам.

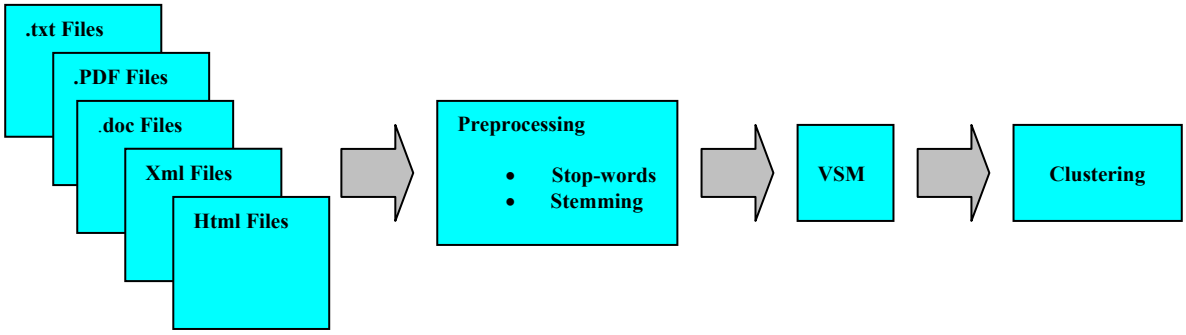

Рис.1. Обработка документов

Описание метода

В системе на первом этапе документы проходят предварительную обработку - сокращение текста для более точной классификации. В нашем случае препроцессинг (предварительная обработка) состоит из двух этапов. Поступивший документ предварительно обрабатывается, перед тем как пройти остальные этапы. Сначала удаляются все стоп-слова из документа. Стоп – слова это набор артиклей, таких как: the, a, in, of и т.д. Затем производится стемминг – это процесс выделения основы слова. Мы решили использовать стемминг, так как он позволяет максимально сократить время обработки документа в системе, что соответственно ведет к оптимизации системы, улучшению ее качества работы. Мы используем стемминг алгоритм Портера.[5] Общая структура этой модели данных начинается с представления любого документа как вектора слов, которые появляются в документах набора данных. Вес (обычно частоты термов) слов также содержится в каждом векторе. После предварительной обработки мы используем векторную модель. Существует большое количество разновидностей векторной модели. Мы используем стандартную векторную модель (1).

$$d_{tfidf} = tf_1 log(n/df_1), tf_2 log(n/df_2), ..., tf_n log(n/df_n) \quad (1)$$

В стандартной векторной модели [6], полагается, что каждый документ является вектором в пространстве термов. В его стандартной форме каждый документ представляется вектором частоты (TF) dtf dtf=(tf1,tf2,.,tfn), где tfi - частота i -й сроки в документе. Обратная частота (IDF) документа в частоте документов i на log (n/dfi), где n - полное число документов в выборке, и dfi - число документов, которые содержат i-ый терм (т.е., частота документа). Наконец, чтобы посчитать длину каждого документа, необходимо каждый вектор документа нормализовать от 0 до 1т.е., $\| d_{tfidf} \|_2 = 1$.

Схожесть документов считается различными методами на основании двух соответствующих по свойствам векторах, например , Jaccard measure, и Euclidean distance [7]. В своей работе мы использовали cosine measure. Каждый документ представляется как вектор частот слов. Длиной векторов мы назовем N. Только частоты N наиболее часто встречаемых слов будут сохранены. Схожесть между двумя документами измеряется как косинус угла между двумя векторами с наиболее схожими документами (2). Чем менше угол, тем больше значение косинуса и на оборот, у менее схожих документов, чем больше угол, тем меньше значение косинуса. Значение косинуса между двумя векторами считается как:

$$cos(X,Y) = \frac{\sum_{i=1}^{n} X_{i,} Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2} \sqrt{\sum_{i=1}^{n} Y_i^2}} \quad (2)$$

В данной формуле (2) X и Y векторы документов.

Следующим этапом в нашем подходе идет иерархическая кластеризация. Суть иерархической кластеризации состоит в последовательном объединении меньших кластеров в большие или разделении больших кластеров на меньшие. Иерархические алгоритмы связаны с построением *дендрограмм* (от греческого dendron - "дерево"), которые являются результатом иерархического кластерного анализа.

Дендрограмма описывает близость отдельных точек и кластеров друг к другу, представляет в графическом виде последовательность объединения (разделения) кластеров. Мы использовали метод Дендрограммы для определения точного количества кластеров (папок) на основе Когда каждый объект представляет собой отдельный кластер, расстояния между этими объектами определяются выбранной мерой.[8] Возникает следующий вопрос - как определить расстояния между кластерами? Существуют различные правила, называемые методами объединения или связи для двух кластеров. Был использован Метод Варда (Ward's method). В качестве расстояния между кластерами берется прирост суммы квадратов расстояний объектов до центров кластеров, получаемый в результате их объединения (Ward, 1963). В отличие от других методов кластерного анализа для оценки расстояний между кластерами, здесь используются методы дисперсионного анализа. На каждом шаге алгоритма объединяются такие два кластера, которые приводят к минимальному увеличению целевой функции, т.е. внутригрупповой суммы квадратов. Этот метод направлен на объединение близко расположенных кластеров и "стремится" создавать кластеры малого размера. [9]Например, на компьютере находится большое количество папок, которые имеют неоднозначные названия («мое», «последние важные документы», и т.д.) Методом дендрограммы определяются папки, которые будут иметь схожие тематические названия.Следующим этапом является метод К-средних [10]. Общая идея алгоритма: задается фиксированное число k кластеров и наблюдения сопоставляются кластерам так, что средние в кластере (для всех переменных) максимально возможно отличаются друг от друга. Данный метод используется для того, чтобы раскластеризировать все неструктурированные (выбранные пользователем) папки с документами по заранее определенным кластерам (папкам), созданные на основе дендрограммы. Завершающим этапом является повторное использование дендрограммы. Внутри сформированной папки на основе дендрограммы создается иерархическая структура вложенных папок.

Заключение

В данной статье использовалась векторная модель для классификации и кластеризации документов. Предложенный подход был применен для текстового репозитория, где использовалась итеративная кластеризация, разделяющая на подкластера каждый класс и заключительным шагом являлась кластерная дендрограмма, для получения иерархической структуры целого набора данных (папок и документов). Интегрированный подход текстовой кластеризации для неструктурированных документов позволяет решать проблемы беспорядочного хранения информации на компьютере и сокращает затраты времени пользователя.

Литература:

[1] «Автоматическая классификация текстов», Юрий Лифшиц, 2006 URL http://logic.pdmi.ras.ru/~yura/internet/06ia.pdf
[2] Bellot, P., & El-Beze, M. (1999).- A clustering method for information retrieval (Technical Report IR-0199). Laboratoire d'Informatique d'Avignon, France.
[3] Zoubin Ghahramani, Gatsby Computational Neuroscience Unit University College London, UK, 2004 URL http://learning.eng.cam.ac.uk/zoubin/course05/ul.pdf
[4] David D. Lewis and Marc Ringuette. A comparison of two learning algorithms for text categorization. In Third Annual Symposium on Document Analysis and Information Retrieval, pages 81–93, 1994. URL http://www.research.att.com/~lewis/papers/lewis94b.ps.
[5] M. Porter. An Algorithm for Suffix Stripping.Program, 14(3):130–137, 1980.
[6] Bellot, P., & El-Beze, M. (1999). A clustering method for information retrieval (Technical Report IR-0199). Laboratoire d'Informatique d'Avignon, France.
[7] The Curse of Dimensionality, URL http://www.stanford.edu/class/cs345a/slides/clustering1.ppt#277,5,
[8] Everitt B. "Cluster Analysis", 2nd edn. Wiley, NewYork, 1993, 283 p.
[9] «Методы кластерного анализа. Иерархические методы» , И.А. Чубукова URL http://www.intuit.ru/department/database/datamining/13/2.html
[10] Bradley, P. S., Bennett, K. P., & Demiriz, A. (2000).Constrained k-means clustering (Technical Report
[11] MSR-TR-2000-65). Microsoft Research, Redmond,WA.