УДК 519.7:007.52

КОМБИНИРОВАННЫЙ ПОДХОД К КЛАССИФИКАЦИИ ТЕКСТОВ

Т.Б. Шатовская[1], И.В. Каменева[2], Ю.А. Тарасов[3]

[1]ХНУРЭ, г. Харьков, Украина, tanita_uk@mail.ru
[2]ХНУРЭ, г. Харьков, Украина, iriska@vk.kh.ua
[3]ХНУРЭ, г. Харьков, Украина, iceman.yt@gmail.com

В данной статье мы представляем сравнительный анализ методов и интегрированный иерархический подход классификации текста основанный на интегрированном подходе используемых дендрограммы и *k*-средних кластеризации. Этот подход позволяет нам представить интегрированный новый метод иерархической кластеризации, который может классифицировать данные без предварительного задания количества классов.

КЛАСТЕРИЗАЦИЯ, ТЕКСТОВАЯ КЛАССИФИКАЦИЯ, АЛГОРИТМ *K*-СРЕДНИХ, МЕТОД ДЕНДРОГРАММЫ, МЕРА СХОЖЕСТИ, ВЕКТОРНАЯ МОДЕЛЬ, СТЕММИНГ, СТОП-СЛОВА

Введение

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, в персональных новостях. В большинстве систем классификации используется традиционный интегрированный подход на основе двух методов: *k*-средних и байесовского классификатора. Эти методы работают с учителем (supervised learning) и требуют непосредственного участия экспертов в процессе решения задачи классификации.

1. Постановка задачи

На сегодняшний день одной из наиболее важных социальных проблем в Украине является формальная безработица среди молодых людей. Даже после получения высшего образования молодые профессионалы достаточно редко могут найти работу, которая совпадает с их специальностью. В частности, сложно найти подходящее место даже после окончания технического вуза. В Харьковском национальном университете радиоэлектроники был создан информационный «Центр-Карьера», который повышает внутренние и внешние уровни информации, оказывая помощь молодым людям найти высококвалифицированную работу после окончания вуза. Была создана Информационная система, которая помогает образовательной инфраструктуре поддерживать информационную взаимосвязь между университетом и индустрией (http://rabota.kture.kharkov.ua). Ежедневно в данную информационную систему поступает большое количество информации от компаний. Такие как: вакансии, новости, объявления и так далее. К сожалению, большая часть полученной информации не структурирована, и вакансии, высланные на e-mail — свободного содержания. Высокая популярность этой системы среди компаний позволяет в обязательном порядке ежедневно непосредственно принимать большое количество текстовых документов. Более того, часто составленные описания вакансий не структурированы, и документы приходится относить к определенным классам инженеру, а это достаточно сложно.

Например: «Инженер-электроник со знаниями MSSQL и Ассемблера». В данной системе в основе стандартные методы классификации текста относят набор вакансий к двум или более главам.

Мы предлагаем интегрированный метод текстовой классификации, используя экспертную оценку (рис. 1).

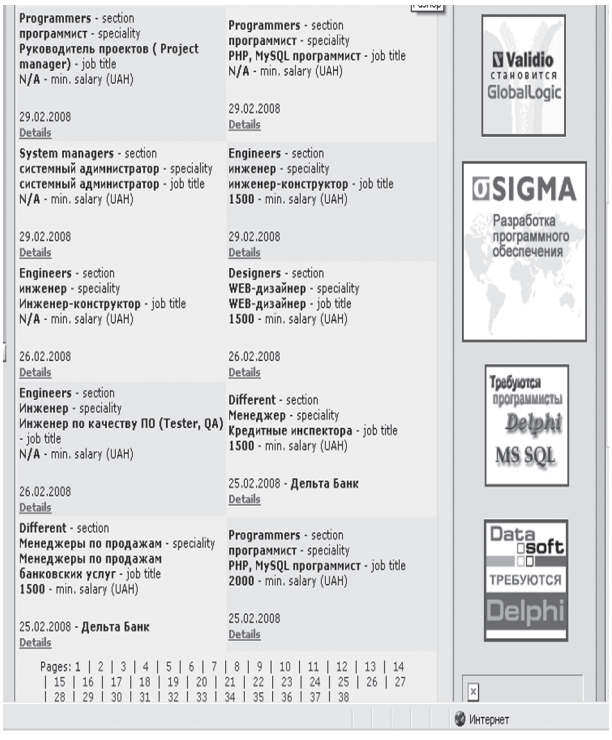

Рис. 1. Набор вакансий

2. Сравнение подходов text Mining

Существует огромное количество классификаторов. Чаще всего применяются Naive Bayesian метод классификации [1, 2], Expectation Maximization (EM) алгоритм [2], Support Vector Machines (SVM) [3] и другие. Все эти подходы обучаются с учителем. В методе обучения с учителем обучающие выборки маркированных документов используются для обучения алгоритма, чтобы построить классификатор. В конечном итоге текстовая классификация определяет документы в одну или более предопределённых категорий. В классификации текста новый текстовый документ назначается в один из уже существующих наборов документов класса.

Текстовая классификация — это лингвистическая и реляционная технология, используемая с целью анализа доступных документов в результате набора полученных ключевых термов из статей и абстрактов. Результаты могут быть представлены графически с картой, которая обеспечивает краткий обзор кластеров и указание взаимосвязи среди них. Главная идея — найти документы, которые имеют много общих слов, и отнести документы с большим количеством общих слов в схожие группы. Текстовая классификация – это автоматическая организация документов в пределах кластера, которые имеют высокое сходство в сравнении друг с другом, но очень несходны с документами в других кластерах [4].

С помощью автоматической классификации мы даем возможность машине решить, к какой из встроенных категории принадлежит текст. В кластеризации машина решает, как предоставленный текст должен быть разделен. Классификация применяется тогда, когда новые тексты классифицируются согласно известной классификации, а кластеризация применяется тогда, когда обнаруживаются новые заранее неизвестные структуры.

3. Иерархическая кластеризация и неиерархическая кластеризация

Иерархическая кластеризация [6] — процесс организации данных в древовидную структуру, основанной на их сходстве. Этот метод очень мощный и полезный для анализа больших наборов данных. Основная идея — создать набор элементов в дереве. Дерево имеет много ветвей. Если элементы подобны друг другу, к ним присоединяются короткие ветви, и наоборот, если их схожесть уменьшается, тогда увеличиваются ветви.

Задача кластеризации текстов состоит в следующем. Имеется некоторое количество текстов. Необходимо сгруппировать эти тексты в соответствии со схожестью их стилей. Такая группировка может быть как одноуровневой («плоской», с выделением таких кластеров, что каждый объект в них является одним из текстов, представленных в наборе, кластеризацию), так и иерархической, когда кластеры, полученные в результате объединения наиболее похожих текстов, сами могут объединяться в кластеры, а кластеры кластеров – в другие кластеры и так далее. Принадлежность текста к некоторому кластеру на определенном уровне иерархической кластеризации может быть однозначной (каждый рассматриваемый текст принадлежит только одному кластеру), или неоднозначной (каждый рассматриваемый текст может принадлежать нескольким кластерам).

Кластеризация документов была использована, чтобы автоматически генерировать иерархические кластеры документов [7].

Результат иерархической агламеративной кластеризации может быть графически представлен как дерево под названием дендрограмма. Алгоритм агламеративной кластеризации [8] порождает кластеры, последовательно соединяя операции. Процесс агломерации начинается с инициализации каждого вектора данных как его собственного кластера. Два кластера соединяются на каждом шаге, и процесс повторяется до тех пор, пока не будет получено желаемое количество кластеров. Если мы рассматриваем общую проблему класетризации, то существует много различных правил слияния. Различные правила слияния предоставляют различные кластерные решения, и таким образом стратегии принимают различные кластерные формы. Метод single linkage (SL) определяет, что кластерная пара соединяется, основываясь на двух самых близких векторах. Метод complete linkage (CL) определяет, что кластер, который соединен, основан на двух ближайших векторах. Вообще, как single-, так и complete-link подходы не очень хорошо работают, потому что они также основываются на собственных решениях, а следовательно и на ограниченном количестве информации (single-link), или они предполагают, что все документы в кластере подобны друг другу (complete-link). Агламеративные алгоритмы всегда детерминированно генерируют схожую кластерную иерархию (таблица).

Сравнение HAC алгоритмов

Method	Combination similarity	Time compl	Optimal	Comment
Single-link	Max sim of any two docs	$O(N2)$	Yes	Chaining effect
Complete-link	Min sim of any two docs	$O(N2 \log N)$	No	Sensitive to outliers

Неиерархическая кластеризация — это процесс монотонно возрастающего ранжирования, поскольку становятся членами больших кластеров. Эти кластерные методы не обладают структурами, подобными дереву, и новые кластеры формируются в последовательной кластеризации или слиянием и разделением кластеров.

Одним из неиерархических методов кластеризации являются методы разбиения[4]. Рассматривается набор кластеров как объектных методов и методов разбиения объектов, чтобы получить требуемые кластеры. В отличие от иерархического метода, этот метод разбиения позволяет объектам изменять группу с помощью процесса кластерного образования. Метод разбиения обычно начинается с начального решения, после которого перераспределение происходит согласно некоторому оптимальному критерию.

Параметр k определяет пользователь, поэтому лучше запустить алгоритмы несколько раз, чтобы выбрать самый лучший параметр k. Также возможно генерировать значение k автоматически, а затем выбирать лучший по проверенным критериям. Более популярный метод разбиения — алгоритм k-средних. Алгоритм k-средних — один из самых простых методов обучающих алгоритмов, который обучается без учителя. Алгоритм k-средних имеет входной параметр, k, и делит набор n объектов в k кластерах таким образом, что результирующее внутрикластерное сходство высоко, но межкластерное сходство низкое. Сначала метод случайным образом выбирает k объектов, каждый из которых изначально представляет центроиды. Каждый оставшийся объект приписывается к кластеру, к которому он является самым ближним, основоваясь на расстоянии между объектом и цетроидом. Затем считается новый центроид для каждого кластера. Этот процесс повторяют, пока оценочная функция не сводится в одну точку [5].

4. Методы текстовой кластеризации

На сегодняшний день векторная модель является широко используемой моделью представления данных для классификации и кластеризации документов. Общая структура этой модели данных начинается с представления любого документа как вектор слов, которые появляются в документах набора данных. Вес (обычно частоты термов) слов также содержится в каждом векторе. Схожесть между двумя документами считается на основании двух соответствующих по свойствам векторов, например, Jaccard measure, и Euclidean distance [9]. Мы использовали cosine measure. Когда описание вакансий получено нашей вэб системой, мы используем метод предварительной обработки. Предварительная обработка — это сокращение текста для более точной классификации. С обработкой методов различные документы могут быть созданы как структурированные представления документа [10]. Обычно задачи предварительной обработки действий включают стандартизацию документа, токенизацию, лематизацию и стемминг [8]. Технология этого процесса рассмотрена на рис. 2.

Мы используем stop-list для исключения стоп-слов. Stop-list — это словарь стоп-слов, который имеет низкую частоту (a, the, and, и так далее) для документов. Стоп-слова всегда удаляют из документов перед преобразованием к векторной модели. Мы используем готовый stop-list. В следующем шаге мы осуществляем stemming алгоритм. Стемминг — это алгоритм, который помогает нам определить значимые части слов. Стемминг удаляет из слов суффиксы рекурсивно. Процесс имеет две цели. В терминах эффективности стемминг сокращает число уникальных слов в индексе, который в свою очередь сокращает пространство памяти, требуемое для индекса, и развивает скорость поискового процесса. В терминах эффективности стемминг улучшается, сводя слова к базовой форме. В нашей работе мы используем стемминг-алгоритм Портера и готовый словарь. Это помогает сократить словарь так, чтобы увеличить скорость нашего метода и улучшить качество алгоритма. На третьем шаге каждый документ использует новую векторную модель tf.idf-exprt. В стандартной векторной модели [10], полагается, что каждый документ является вектором в пространстве термов. В его стандартной форме каждый документ представляется вектором частоты (TF) $dtf\ dtf = (tf_1, tf_2, ., tf_n)$, где tf_i — частота i-й сроки в документе. Обратная частота (IDF) документа в частоте документов i на $\log (n/df_i)$, где n — полное число документов в выборке, и df_i — число документов, которые содержат

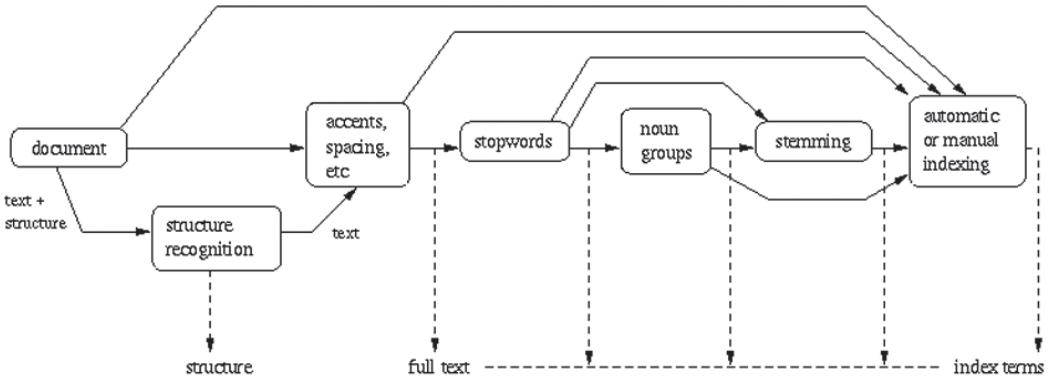

Рис. 2. Последовательность предварительной обработки документа

i-ый терм (то есть, частота документа). Наконец, чтобы посчитать длину каждого документа, необходимо каждый вектор документа нормализовать от 0 до 1, то есть $\| d_{tfidf} \|_2 = 1$. Мы также добавляем экспертную оценку для каждого терма, потому что для вакансий очень важно оценить значение каждого терма в документе. Мы использовали *tf* × *idf-exprt*(*f*) как

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

Все параметры остаются без изменения, но f — это экспертная оценка, где $f = 0...1$. Для подсчета схожести между документами часто применяется cosine measure, которое определяется как

$$\cos ine(d_i, d_j) = \frac{\langle d_i" \, d_j \rangle}{\|d_j\|_2 \times \|d_i\|_2} = \frac{\sum_{t=1}^{t} d_i \times d_j}{\sqrt{\sum_{t=1}^{t} d_i^2} \times \sqrt{\sum_{t=1}^{t} d_i^2}},$$

где d_i и d_j — компоненты векторных документов; t — размерность вектора.

Расстояние между векторами d_1 и d_2 представляется как косинус угла между ними (рис. 3).

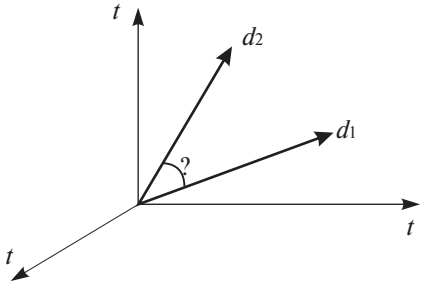

Рис. 3. Схожесть между d_1 и d_2

Мера расстояния для векторов удовлетворяется следующими свойствами:

– если два вектора совпадают целиком их сходство должно быть максимально, то есть равняться 1;

– если два вектора не имеют никаких ключевых общих слов, то есть, если вектор запроса не имел позитивных весов документа, вектор имеет вес 0, и наоборот – или другими словами, если векторы ортогональны – сходство должно быть минимально, то есть равняться 0.

Во всех других случаях схожесть должна быть от 1 до 0. Документы, которые соединены друг с другом в векторном пространстве показывают схожесть между ними, где *x* и *y* это cosine measure (схожесть) между документами (рис. 4).

На четвертой стадии кластеризации, на базе Ward's linkage правила и cosine measure, мы конструируем дендрограмму, чтобы получить начальное разделение входных документов. Так как мы не знаем предварительного количества классов в наших данных, первая дендрограмма показывает глобальное число классов.

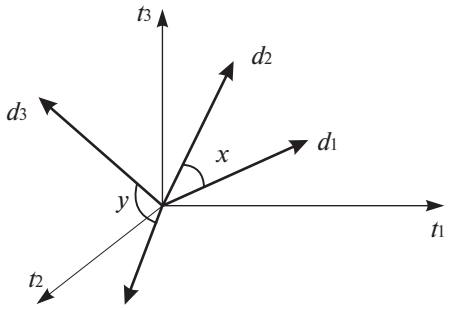

Рис. 4. Схожесть между документами

После этого мы применяем метод кластеризации *K*-средних к каждому классу дендрограммы. Для оптимальных результатов мы использовали меру компактности кластеров. После этого кластеризация дендрограммы будет применена к каждому подкласетру и общее количество будет оценено. Как заключительный результат мы получаем кластеры и все их подкластеры. Этот подход применяется непосредственно к вакансиям, группирующимся среди 20 категорий, и частота появления ошибок составила 8% из-за некорректного группирования.

Заключение

Комбинированный подход использования итеративной кластеризации и агломеративных методов используются для кластеризации вакансий веб-сайта. Векторная модель использовалась для классификации документа и кластеризации. Главными шагами предоставленного подхода являются: кластерная дендрограмма для целого текстового репозитория, затем итеративная кластеризация, которая используется для разделения на подкластеры внутри каждого класса. и заключительным шагом является кластерная дендрограмма для получения иерархической структуры целого набора данных.

Список литературы: http://en.wikipedia.org/wiki/Naive_Bayes_classifier 2. http://apex.sjtu.edu.cn/apex_wiki/dwyak?action=AttachFile&do=get&target=aaai07.pdf3.http://jmlr.csail.mit.edu/papers/volume2/tong01a/tong01a.pdf4.http://www.cs.sfu.ca/~ester/papers/Encyclopedia.pdf5.http://www.nada.kth.se/~rosell/undervisning/sprakt/irintro060801.pdf 6.http://www.ims.uni-stuttgart.de/lehre/teaching/2007-SS/ir/hier/hier.pdf 7. Daphe Koller and Mehran Sahami, Hierarchically classifying documents using very few words, Proceedings of the 14th International Conference on Machine Learning (ML), Nashville, Tennessee, July 1997, Pages 170-178. 8. J. Boberg and T. Salakoski. General formulation and evaluation of agglomerative clustering methods with metric and non-metric distances. Pattern Recognition, 26(9):1395–1406, September 1993. 9. http://www.ee.ucr.edu/~barth/EE242/clustering_survey.pdf 10. [Salton et al., 1975] G. Salton, A. Wong, and C.S. Yang. A vector space model for information retrieval. Journal of the American Society for Information Science, 18(11):613-620, Nov.1975.

Поступила в редколлегию 28.04.2008