Система отбора вакансий для разработчиков программного обеспечения

Д.Ю. Ходаков[1], *К.А.Терехина*[1,2]

[1]Московский физико-технический институт (государственный университет)

[2]Российская академия народного хозяйства и государственной службы при Президенте РФ

Введение

На сегодняшний день существует большое разнообразие сайтов по подбору вакансий для разработчиков программного обеспечения. Однако даже самые современные сайты по подбору персонала не всегда удачно справляются со своей задачей. Обычно подбор вакансий основывается на ожидаемой заработной плате без проведения обработки естественного языка. Обработка естественного языка дает возможность для соискателей найти работу, подходящую под их навыки, а также оценить возможную максимальную и минимальную заработную плату.

Обзор работ

Многочисленные недавние исследования посвящены обработке естественного языка, некоторые из таких работ представлены в хорошо известных монографиях [3], [4]. В данной работе мы делаем акцент на обработку естественного языка в коротких текстах. Один из хороших примеров обработки коротких текстов является работа Александрова и др. [1], где проводилась кластеризация абстрактов статей. В работе Раманатхана и др. [7], было предложено использовать ключевые слова с дополнениями из Википедии, чтобы группировать популярные новости (RSS/Atom). Однако следует отметить, что кластеризация коротких текстов сложнее, чем длинных текстов. Одна из таких трудностей это работа с отбором ключевых слов в коротких тестах и расчет матрицы расстояния, что подробно описано в работе [2].

Урвой и др. [9] изучали другие проблемы: меру близости и алгоритм определения скрытого схожего стиля для веб-спама. В их работе они описывали популярную на сегодняшний день меру Жаккара и меру Диса. Тем не менее, расчет близость между текстами эффективно при малом количестве текстов, при большом же количестве происходит квадратичное увеличение затрат вычислительной мощности, что неприемлемо в масштабе веб поисковика.

Существует большое разнообразие алгоритмов кластеризации[5]. Самым популярным является метод k-средних. Сравнение методов k-медоид, PAM и CLARANS, показывает, что PAM и CLARANS показывают одинаковые результаты [8]. Алгоритм PAM является самой распространенной реализацией кластеризации k-медоид. Он может быть использован с любой мерой близости. Алгоритм k-средних может не сойтись, так как он может быть использован только с расстоянием, которое согласуется со средним. Алгоритм PAM требует больших вычислительных затрат и работает дольше чем алгоритм k-средних.

Поч и другие [10] недавно представили систему, которая предлагает ранжированных список подходящих вакансий для соискателей работы, основываясь на наборе навыков кандидата, вакансий и пар сочетаний вакансия - работник.

Матрица близости

В данной работе мы использовали коэффициент Жаккара для расчета матрицы близости. Коэффициент Жаррака это статистическая мера, используемая в обработке естественных текстов для сравнения схожести пары документов. Она определяется как мощность пересечения, деленная на мощность объединения набора текстов. Пусть d_i и d_j два рассматриваемых документа, тогда коэффициент Жаккара будет выглядеть следующим образом:

$$Jaccard(d_i, d_j) = \frac{|d_i \cap d_j|}{|d_i \cup d_j|}$$

Эксперименты

Исходные данные в работе представляют собой предложения о работе для разработчиков программного обеспечения, размещенные на сайте hh.ru. Ежедневно с сентября 2014 по май 2015 года данные о вакансиях скачивались и загружались в базу данных в автоматическом режиме. Было собрано 65 000 вакансий, из них 5 700 уникальные, которые в дальнейшем и использовались в исследовании. Все предложения о работы были сохранены в html формате. Чтобы извлечь информацию была проведена обработка всех html страниц. Для этого существует библиотеки: BeautifulSoup, html5lib, lxml.etree.HTML, mochiweb html и т.д. В работе использовалась библиотека BeautifulSoup для Python 3 из соображений удобства авторов. Каждое объявление о работе было представлено в таблице, которая содержала информацию об опыте работы, величине

заработной платы и присутствие или отсутствие каждого из 107 навыков (python, c, c++, java, mysql и др).

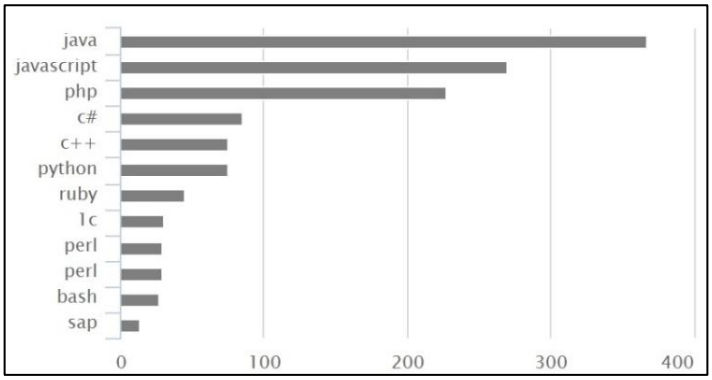

Рис. 1. Количество вакансий по языкам

В качестве данных для кластеризации использовался вектор из нулей и единиц, который соответствовал навыкам вышеперечисленным навыкам. В качестве метода кластеризации был выбран метод k-медоид. Для расчета матрицы дистанций была использована мера Жаккара. Однако, чтобы посчитать попарную близость между 5700 вакансиями надо обработать матрицу 5700*5700, что было бы достаточно затратно. Для сокращения времени расчётов мы использовали тот факт, что матрица симметричная, а так же технику расчета дистанций Жаккара, основанную на BitMap. Вектор из нулей и единиц был преобразован в целое число, что позволило поместить все данные кэш процессора. Алгоритм Дуда-Харта [6] выделения кластеров показал наличие 10 ярко выраженных кластеров.

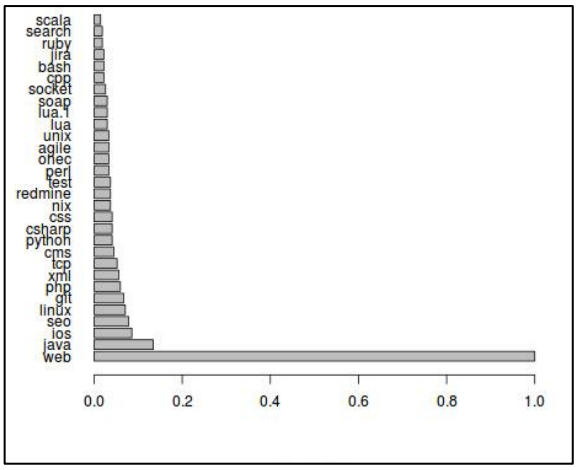

Рис. 2. Кластер #4

Согласно экспертной оценке кластер 4 на рисунке 2 является кластером веб-разработчиков. Из графика видно, что присутствуют навыки с большей и меньше вероятностью принадлежащие этому кластеру. Аналогичные результаты были получены и для остальных кластеров.

Литература

[1] M. Alexandrov, A. Gelbukh1, and P. Rosso,*An Approach to Clustering Abstracts*, Natural Language Processing and Information Systems, Lecture Notes in Computer Science Volume 3513, 2005, pp 275-285

[2] D. Pinto, *On clustering and evaluation of narrow domain short-text corpora*, Doctoral Dissertation, Polytechnic University of Valencia, Spain, 2008

[3] C. Manning,P. Raghavan,H. Schutze, *Introduction to Information Retrieval*, Cambridge University Press, 2008

[4] R. Baeza-Yates,B. Ribero-Neto, *Modern Inform. Retrieval*, Addison Wesley, 1999

[5] J. Hartigan, *Clustering Algorithms*, Wiley, 1975

[6] R. Duda, and P. Hart, *Pattern Classification and Scene Analysis*, Wiley, 1973

[7] S. Banerjee, K. Ramanathan, A. Gupta, *Clustering Short Texts using Wikipedia*,HP Laboratories, India, 2008

[8] A. Reynolds, G. Richards, B. Iglesia and V. Rayward-Smith, *Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms*, Journal of Mathematical Modelling and Algorithms (2006) 5: 475504, DOI: 10.1007/s10852-005-9022-1

[9] T. Urvoy, T. Lavergne, P. Filoche, *Tracking Web Spam with Hidden Style Similarity*,France Telecom RD, 2006

[10] M. Poch, N. Bel1, S. Espeja,*Ranking Job Offers for Candidates: learning hidden knowledge from Big Data*, LREC 2014: 2076-2082

[11] Y. Fang, J. Wang, *Selection of the number of clusters via the bootstrap method*, Computational Statistics and Data Analysis 56 (2012) 468477