INFORMATYKA EKONOMICZNA BUSINESS INFORMATICS 4(30) · 2013
ISSN 1507-3858

Karolina Kuligowska, Mirosława Lasek
University of Warsaw
e-mail: {kkuligowska; mlasek}@wne.uw.edu.pl

TEXT MINING IN PRACTICE: EXPLORING PATTERNS IN TEXT COLLECTIONS OF REMOTE WORK JOB OFFERS

Summary: The aim of this paper is to give an insight into text mining techniques in the context of unstructured text collections of location independent job offers. In order to extract useful information, uncover interesting patterns and features of remote work, we analyze the five most popular and most visited websites containing job offers. We examine clusters of remote job offers, the keywords describing those clusters, as well as the linkages between strongly associated terms describing mobile work offers. It is interesting to observe the maturity of the text mining tools which have broadened their applications to new research topics and have become suitable to explore new phenomena.

Keywords: text mining, text analytics, clustering, concept linking, remote work, telecommuting.

1. Introduction

Since most human knowledge is maintained in a textual form [Lin 2001], thus text analytics methods and techniques are constantly being developed, and this development has accelerated recently [Agrawal, Batra 2013; Mahesh, Suresh, Vinayababu 2010; Patel, Soni 2012; Ramanathan, Meyyappan 2013]. Previously focused on calculations, nowadays computer power more and more caters for text issues such as computational linguistics, natural language processing and text mining. Word patterns, context recognition, and term linkages constitute the subject of insightful analysis. With emerging text mining tools, scientists can discover interesting results in many fields by exploring vast amount of text data.

The era of innovative technology also influences the way of working. It is not necessarily uncommon nowadays to use the Internet connection for working instead of going to the real physical office. Remote working has evolved quickly and freelancing has become an increasingly dynamically developing field. Nevertheless, there is a widespread belief that telework job offers are focused mainly on those who know computer graphics and have skills in programming. Is this a correct conviction?

The aim of this paper is to explore unstructured text of remote work job offers by applying text mining techniques. The paper is organized as follows. Section 2 presents data source and software applied in our analysis. Section 3 introduces studies about location independent work and reviews in brief remote work features. Section 4 combines the discussed issues in one analytical frame, presents the obtained results and describes the conclusions of tables and figures. Section 5 provides a summary of our findings. Finally, the conclusions arising from this paper are considered in Section 6.

2. Data source and software applied

The remote job offers examined in our text mining analysis originate from the following websites: Careerbuilder[1], Remoteemployment[2], Monster[3], Jobamatic[4], Simplyhired[5]. These international websites are some of the world's most widely recognized portals with job offers for remote and regular workers. However, there are many more job portals in the Internet, beyond the websites used in our analysis.

With regard to the software tool, we performed text parsing, clustering and concept linking techniques using SAS Text Miner 4.2 software within the SAS Enterprise Miner 6.2 environment. By executing macro % tmfilter, we extracted text of job offers from each website. Subsequently, each set of unstructured text collection was turned into an adequate structured table and analyzed within the SAS business analytics flexible framework.

3. Remote work

3.1. Telework features

Remote work, as an innovative form of employment, is perceived as an alternative for the traditional in-office work environment. In the last decade, companies have slowly shifted towards a virtual workplace, and employees have quickly adapted to fulfill this business demand. As technological advances have provided mobile electronic media of communication, companies have realized the benefits of the virtual workplace trends and flexible work arrangements [Busch, Nash, Bell 2011; Lister, Harnish 2011]. A survey conducted in 2011 by the global research company Ipsos, revealed that telecommuting is primarily taking place in the emerging markets of the Middle East, Africa, Latin America, and Asia-Pacific [Gottfried 2012]. Moreover, according to the Forrester Research forecasts, telework will apply to 43% of US

[1] http://www.careerbuilder.com/Jobs/Keyword/Remote/ [accessed: April 2013].
[2] http://www.remoteemployment.com [accessed: April 2013].
[3] http://jobsearch.monster.com/search/?q=remote [accessed: April 2013].
[4] http://momstowork.jobamatic.com/a/jobs/find-jobs/q-Remote [accessed: April 2013].
[5] http://www.simplyhired.com/a/jobs/list/q-remote/fjt-telecommute [accessed: April 2013].

workers by 2016 [Schadler, Brown, Burnes 2009]. Remote work is practiced globally and there is no doubt that the telecommuting trend is steadily growing.

Mobile work executed at a distance is described by numerous terms and synonyms such as: remote jobs, telework, telecommuting jobs, online jobs, home based work, flexible jobs and location independent work. Taxonomy researchers distinguish three main forms of telework: fixed-site telework, mobile telework, and flexi-work [Garrett, Danziger 2007]. In a broad sense, a remote location worker is an employee who performs his or her work outside the workplace, communicating the results of his or her work by means of electronic communication, thus eliminating distance restrictions or any other problems associated with traditional commuting practices [Watad, Jenkins 2010; Fuhr, Pociask 2011]. This type of work is quite well suited for freelancers such as graphic designers, computer programmers and interpreters/translators. Furthermore, especially in the U.S., remote workers offer virtual assistant services, i.e. secretarial services, customer services and sales executed over the phone.

3.2. Telework research

As the phenomenon of telework is constantly spreading around the world, there are more and more research studies done on this subject. They mainly focus on measuring: the effectiveness and productivity of remote work [Bloom et al. 2013; Dutcher, Saral 2012; Teh, Ong, Loh 2011], the activation of people at risk of unemployment or the socially excluded [Baker, Moon, Ward 2006; Stroińska 2012], and the psychological impact of teleworking on human stress, emotions and health [Mann, Holdsworth 2003; Ward, Shabha 2001]. Nonetheless, there is little cross-sectional research concerning the features of telework accessible for freelancers willing to work remotely. Therefore we will conduct our text mining study to uncover descriptive keywords, explore patterns and common features that link various remote job offers. This approach adds a new research area to widely known standard text mining application fields such as bioinformatics, business intelligence and customer relationship management [Gupta, Lehal 2009; Jusoh, Alfawareh 2012].

4. Text mining patterns exploration

We started our analysis by making five independent text mining models executed separately for each dataset. First, we extracted the text of job offers from each website and turned it into an adequate input dataset. Then we extracted parts of speech and gathered structured information from unstructured text, in search of relevant terms hidden in the text database [Blansché et al. 2010; Kaur, Aggarwal 2013]. For this purpose we subsequently performed three phases: text parsing, text filtering and text mining, using the appropriate nodes, as shown in Figure 1.

The Text Parsing node gathered the statistical data about the terms, such as the number of terms, number of documents, and term frequencies in each dataset. In the text parsing process we decided to focus only on one part of speech, namely adjectives. Adjectives describe nouns in terms of many qualities and therefore they constitute relevant elements giving more information than any other part of speech, especially in the case of remote work offers. The Text Parsing node allowed us to modify our output set of parsed terms by limiting our text parsing results to terms with the role of Adjective.

Next, we used the Text Filter node in order to keep only sufficiently important terms and to remove irrelevant terms. Therefore we sorted the results for each set of text data by term weight in a descending mode. Terms with a weight below 0.1. were rejected from our analysis.

Finally, we applied the Text Mining node to examine the information that exists in each dataset. The Text Mining node allowed us to execute categorization, clustering, and keyword extraction, which are the text mining's major tasks [Gharehchopogh, Khalifehlou 2012].

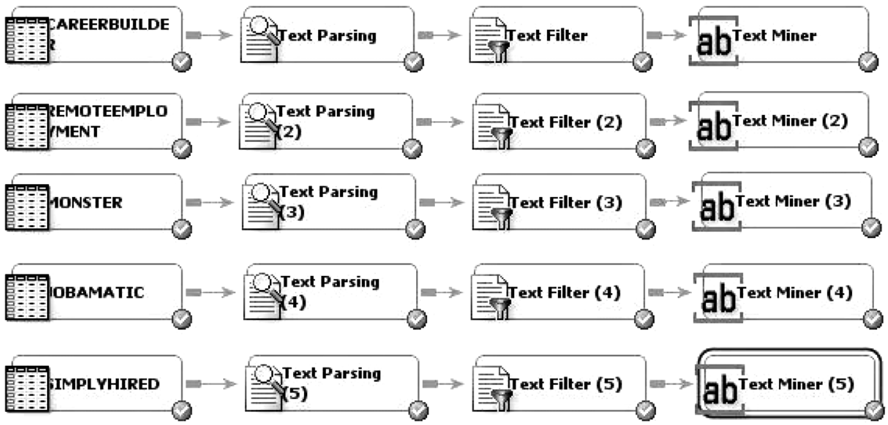


Figure 1. Five text mining models

Source: own elaboration.

Clustering, defined as a "process of partitioning a dataset into clusters, so that elements of the same cluster are more similar to each other than to elements of different clusters" [Su, Kogan, Nicholas 2010], constitutes a fundamental text mining method. It allows to achieve an organized overview of the concepts contained in text documents and improves similar document detection [Bolasco et al. 2005; Vidhya, Aghila 2010]. In order to find groups of similar offers in our text collection, we executed similarity based clustering. We applied the iterative Expectation Maximization algorithm, a clustering method which is most frequently used in a wide variety of applications [Aggarwal, Zhai 2012]. The commonly applied Singular Value Decomposition technique [Radovanović, Ivanović 2008] was used for reducing the number of dimensions. Each cluster was described by 10 keywords, and we also allowed the unclustered outliers to be excluded from analysis.

Ultimately, for the selected adjective with the highest weight, we were eager to identify other terms that correlate the strongest with it. Therefore for each dataset we executed concept linking. This technique is used to find and present highly related terms, whereas the strength of association between these terms is measured by the chi-square statistic. Since the visualization of related expressions helps to derive new insights and novel patterns in text collection [Don et al. 2007; Gupta, Lehal 2009], the linkage of correlated concepts is displayed in the form of a hyperbolic tree graph with nodes that are expanded when necessary. Thus concept linking visualization enriches the traditional searching methods with advanced browsing capabilities.

4.1. Careerbuilder

In the Careerbuilder dataset of 175 remote work job offers, we identified 480 adjectives occurring in the text. The three adjectives with the highest weight were: *liable* (0.927), *security* (0.921) and *advanced* (0.913).

As a result of applying the Expectation-Maximization clustering algorithm, two clusters emerged. Each cluster is described by ten keywords (Descriptive terms), the number of documents in a cluster (Freq), the percentage of documents in a cluster (Percentage), and the root mean squared standard deviation (RMS Std.), as presented in Figure 2. As stated before, the unclustered outliers were excluded from the analysis. Therefore the total number of documents in the clusters (169) is lower than the overall number of analyzed documents in the dataset (175).

Clusters

#	DESCRIPTIVE TERMS	FREQ	PERCENTAGE	RMS STD.
1	other, junior, similar, different, future, able, remote, private, senior, last	100	0.5714285714...	0.1252072...
2	new, more, best, top, next, additional, successful, career, available, temporary	69	0.3942857142...	0.2671486...

Figure 2. Clusters sorted by Frequency.

Source: own elaboration.

The first cluster grouped 100 job offers described by such terms as: *other, junior, similar, different, future, able, remote, private, senior, last*. Most probably these offers refer to high school or college/university students and were aimed at 3rd, 4th or final year students able to work remotely in the private sector. The descriptive terms of the second cluster, grouping 69 job offers, were: *new, more, best, top, next,* *additional, successful, career, available, temporary*. This may indicate a group of temporary or additional part-time jobs which may contribute to a successful career in the future.

For one of the terms with the highest weight, i.e. *advanced*, we executed the concept linking visualization, as presented in Figure 3.

Figure 3. Concept linkages for the term advanced

Source: own elaboration.

On the concept linking hyperbolic tree graph, the term *advanced* is displayed in the center of the structure. We can observe that the term *advanced* is surrounded by the following adjectives: *new*, *best*, and *career*. The thickness of the line between the concepts represents the strength of association, and the thicker line indicates a closer association. Therefore *advanced* is strongly correlated with *new* and *career*, as well as with *successful* and *additional* in the expanded view mode.

4.2. Remoteemployment

In the Remoteemployment database we identified 758 adjectives occurring in the text of 231 remote work job offers. The three adjectives with the highest weight were: *clinical* (0.955), *residential* (0.950) and *industrial* (0.936).

Once again we applied Expectation Maximization as the clustering method, and we obtained two clusters, as presented on Figure 4. The first cluster grouped 115 job offers and was described by quite irrelevant adjectives such as *both, unique, big,* *same, little, much, few, great, important*, and *better*. Whereas the second cluster gathered 113 documents and its descriptive terms related to contract duration, such as *part-time, full-time, interim, temporary*, among others.

Clusters

#	DESCRIPTIVE TERMS	FREQ	PERCENTAGE	RMS STD.
1	both, unique, + big, same, + little, much, + few, + great, important, better	115	0.4978354978...	0.1667515...
2	part-time, full-time, interim, temporary, part, + fast, + large, no, successful, good	113	0.4891774891...	0.1509026...

Figure 4. Clusters sorted by Frequency

Source: own elaboration.

In the case of the Remoteemployment dataset, the clusters look unclear and insignificant for further inference, therefore we continued to examine this dataset with the concept linkage tool. For the term with the highest weight, namely *clinical*, we executed the concept linking hyperbolic tree graph, as you can see in the following Figure 5.

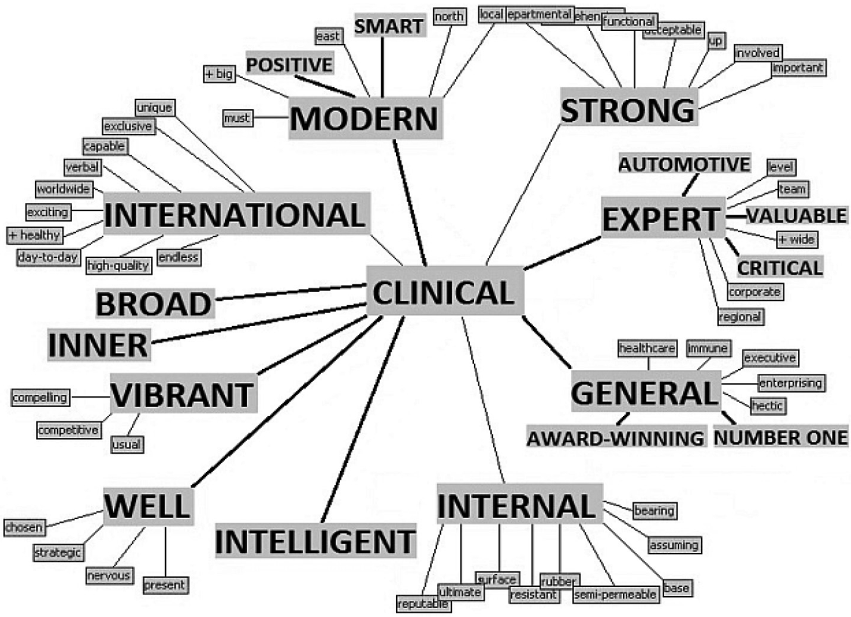


Figure 5. Concept linkages for the term clinical

Source: own elaboration.

The term *clinical* is surrounded by the following adjectives: *international*, *modern, strong, expert, general, internal, intelligent, well, vibrant, inner*, and *broad*. The term *clinical* is highly associated with *modern, expert, general, intelligent, well, vibrant, inner,* and *broad.* Further analysis of expanded nodes of terms *modern, expert* and *general* indicates the close association with adjectives describing the candidate's personal features required in work: *positive, smart, automotive, valuable, critical, award-winning, number one.*

4.3. Monster

In the Monster database we identified 970 adjectives out of the 179 remote work job offers text. Adjectives with the highest weight were: *sole* (0.948), *automatic* (0.921), and *editorial* (0.904).

As a result of the clustering procedure we obtained three clusters, as presented in Figure 6.

Clusters

#	DESCRIPTIVE TERMS	FREQ	PERCENTAGE	RMS STD.
1	+ high, responsible, + large, excellent, + fast, common, other, daily, biweekly, monthly	74	0.4134078212...	0.1475900...
2	additional, human, real, legal, dental, free, retail, administrative, austin, top	63	0.3519553072...	0.0971502...
3	subject, invalid, marital, last, remote, veteran, similar, proud, first, united	18	0.1005586592...	0.0665655...

Figure 6. Clusters sorted by Frequency

Source: own elaboration.

The first cluster contained 74 job offers and included offers characterized by contract duration (*daily, biweekly, monthly*), as well as by candidate's features (*responsible, excellent, fast*). The second cluster gathered 63 offers and was described by terms concerning the type of job, such as *administrative*, *legal, dental, retail* and *additional*. The third cluster, with descriptive terms for 18 offers, turned out to be hard to interpret.

For the term *automatic* we executed a concept linking graph, as shown in Figure 7.

The term *automatic* is surrounded by *leading, ideal, exciting, limited, team, talented, advanced, relevant, electronic, up*, and *normal*. The thick connection lines indicate close associations between the term *automatic* and the adjectives: *leading, ideal, exciting, talented, advanced,* and *relevant*. A further expanded view once again relates to a description of the potential candidate's personal features (*experienced, responsible, all-round, able, hard-working, smart*), and salary period conditions (*weekly, monthly, daily*).

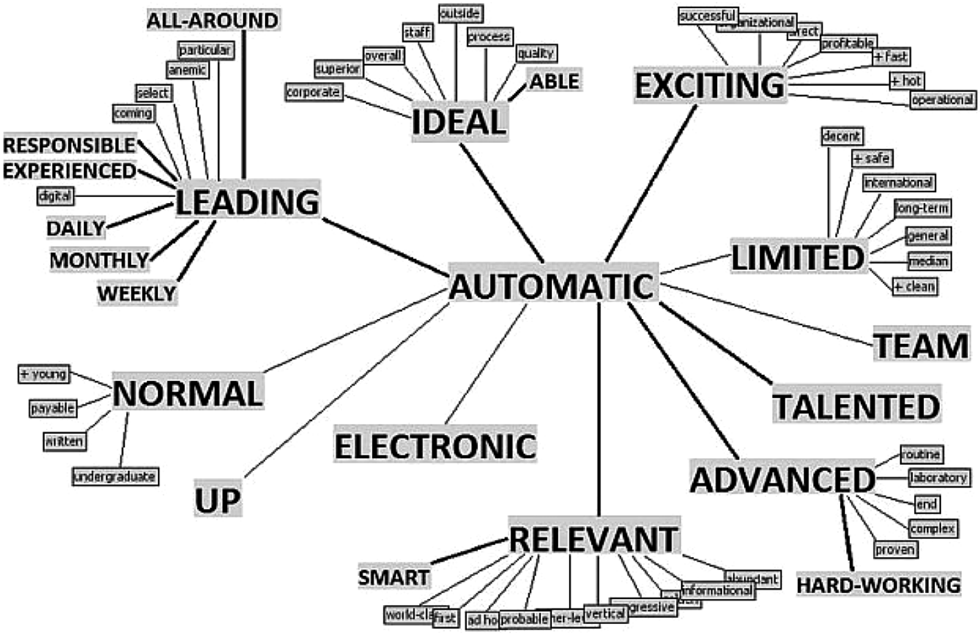


Figure 7. Concept linkages for the term automatic

Source: own elaboration.

4.4. Jobamatic

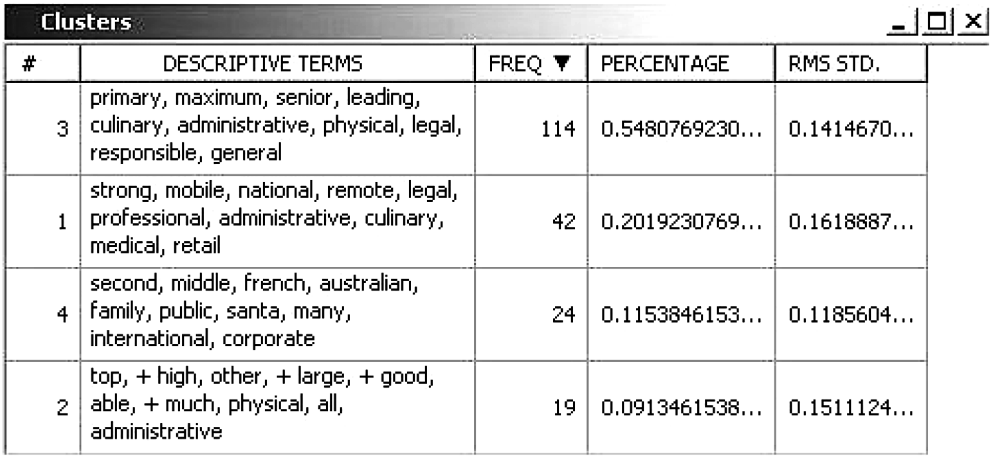
In the Jobamatic database of 208 remote work job offers we identified 778 adjectives occurring in the text. Among the most weighted adjectives were: *magnificent* (0.969), *statistical* (0.906), and *advisory* (0.895).

From the Expectation Maximization clustering algorithm there emerged four clusters, as presented in Figure 8.

The third and first clusters, that grouped 156 job offers altogether, referred to miscellaneous types of job offers (*culinary, administrative, physical, legal, medical, retail*). In the fourth cluster, with descriptive terms for 24 offers, we can find geographical keywords such as French, Australian and international. The keywords of the second cluster, gathering 19 job offers, look vague and it seems hard to find any particular topic or theme for that cluster.

The adjective *statistical* occurred among terms with the highest weight, therefore we executed a concept linking visualization for this term, as presented in Figure 9.

The term *statistical* appears to be related with the adjectives *culinary, administrative, responsible, technical, main, human, leading, social, legal, online,* and *academic*. Since it can be judged by a visual observation of the thickness of the lines, the term *statistical* reveals a higher association with the terms *administrative,*



Clusters

#	DESCRIPTIVE TERMS	FREQ ▼	PERCENTAGE	RMS STD.
3	primary, maximum, senior, leading, culinary, administrative, physical, legal, responsible, general	114	0.5480769230...	0.1414670...
1	strong, mobile, national, remote, legal, professional, administrative, culinary, medical, retail	42	0.2019230769...	0.1618887...
4	second, middle, french, australian, family, public, santa, many, international, corporate	24	0.1153846153...	0.1185604...
2	top, + high, other, + large, + good, able, + much, physical, all, administrative	19	0.0913461538...	0.1511124...

Figure 8. Clusters sorted by Frequency.

Source: own elaboration.

Figure 9. Concept linkages for the term statistical

Source: own elaboration.

responsible, technical, leading, social, *academic* than it exhibits with other terms in the text collection. The expanded sub-nodes are worth a closer look, in particular the term *retail,* related to the type of work, as well as other concepts associated to *statistical* (*financial, forensic*).

4.5. Simplyhired

In the Simplyhired dataset we identified 286 adjectives occurring in the text of 79 remote work job offers. The most weighted adjectives were: *proud* (0.926), *associate* (0.897), and *interactive* (0.885).

In the next step we executed clustering and we obtained four clusters, as presented in Figure 10. The first cluster, in spite of being the largest and gathering over 40% of documents, was unclear and therefore not particularly helpful to uncover any interesting topic in the Simplyhired job offers. Whereas the third and second cluster, grouping altogether 32 job offers, referred to various types of jobs (*commercial, medical, local*), and the way of their performance (*simple, easy, fast*). The fourth cluster referred to the duration of the contract (*monthly, temporary, seasonal*).

Clusters

#	DESCRIPTIVE TERMS	FREQ ▼	PERCENTAGE	RMS STD.
1	remote, clear, next, inaccurate, seeing, + late, similar, relevant, recent, free	32	0.4050632911...	0.1824150...
3	commercial, medical, senior, + full, new, pacific, local, all, inaccurate, seeing	20	0.2531645569...	0.2228919...
2	dream, simple, awesome, + easy, + few, single, good, + fast, important, special	12	0.1518987341...	0.1626406...
4	median, monthly, male, temporary, seasonal, nearby, marital, + little, retail, professional	11	0.1392405063...	0.1158223...

Figure 10. Clusters sorted by Frequency

Source: own elaboration.

Subsequently, we executed a concept linking for one of the terms with the highest weight, namely *interactive*, as presented in Figure 11.

According to the concept linking hyperbolic tree graph, the term *interactive* is associated with the terms: *medical, part-time, regional, direct, exciting, corporate, virtual, elite, acute, young,* and *pivotal*. The width of the line outgoing from the centered adjective reveals the strong correlation with the term *part-time* and its sub-nodes relating to contract duration (*monthly, full-time, seasonal, temporary*) and desirable candidate's characteristics (*self-employed, graphic*).

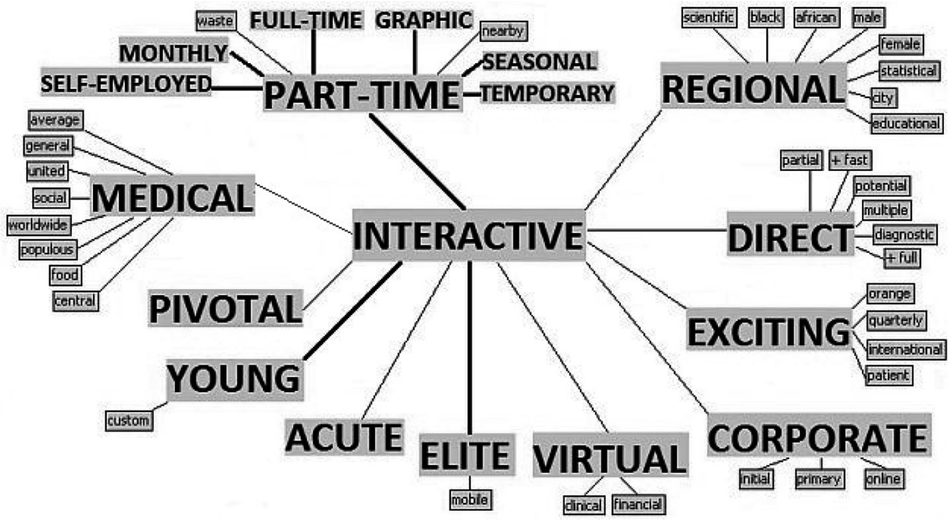


Figure 11. Concept linkages for the term interactive

Source: own elaboration.

5. Discussion

We conducted our text mining study to uncover relevant features pertaining to remote job offers. Initially, we examined adjectives with the highest weights which emerged in each dataset (*liable, security, advanced, clinical, residential, industrial, sole, automatic, editorial, magnificent, statistical, advisory, proud, associate, interactive*). They tend to describe the personal features expected from the candidates. Without specifying a particular profession, they indicate the characteristics of particular importance in any type of work: *liable, advanced, editorial, magnificent, advisory, proud, interactive*.

Subsequently, each of the five datasets was subjected to clustering. Hence, job offers were divided into sets of similar content, described by the ten most relevant keywords. By analyzing them we identified four of the most recurrent attributes of remote work offers. The first attribute related to contract duration, and it covered terms such as *part-time, full-time, interim, temporary, daily, biweekly, monthly, seasonal*. The second group of telework job offers contained names of miscellaneous types of job (*culinary, administrative, physical, legal, medical, retail, dental, commercial, local*), as well as the way of accomplishing them (*simple, easy, fast*). The third topic that appeared in advertisements concerned the candidate's features: *responsible, excellent, fast, junior, able, senior, last [year], best, top, successful, available*. The fourth feature that was distinguished among the clusters, constituted geographical and/or linguistic indicators such as *French, Australian,* and *international*. However, in several cases the clusters turned out to be described by quite irrelevant or meaningless adjectives. It was hard to attribute any particular topic or theme for these clusters, and they were not considered as significant for further analysis.

Afterwards, for each dataset we executed hyperbolic tree graphs in order to visually explore linkages between strongly associated terms describing mobile work offers. The adjectives with the highest weights were surrounded by closely related terms, among which we could distinguish some interesting patterns. The term *advanced* was strongly correlated with *new* and *career*, as well as with *successful* and *additional* in expanded view mode. The term *clinical* was highly associated with *modern, expert, general, intelligent, well, vibrant, inner*, and *broad*, as well as with adjectives describing the candidate's personal features required in work: *positive, smart, automotive, valuable, critical, award-winning, number one*. Close associations were detected between the term *automatic* and the adjectives: *leading, ideal, exciting, talented, advanced*, and *relevant*. The expanded view mode for the term *automatic* related to a description of the desired candidate's characteristics (*experienced, responsible, all-round, able, hard-working, smart*), and conditions of salary period (*weekly, monthly, daily*). The term *statistical* revealed close associations with the terms *administrative, responsible, technical, leading, social* and *academic*. The expanded sub-nodes indicated *retail, financial*, and *forensic* as the type of work associated with statistics. *Interactive* revealed a strong correlation with the term *part-time* and its sub-nodes relating to contract duration (*monthly, full-time, seasonal, temporary*) and to the potential candidate's personal features (*self-employed, graphic*).

6. Conclusions

This paper has presented an analysis of remote work offers by using text mining techniques. The focus has been centered on text parsing, text filtering, and text mining carried out for the needs of standard browsing, clustering, concept linking and further inference. Remote job offers were separated into general clusters and afterwards each cluster was automatically presented by essential descriptive keywords. By analyzing those keywords we identified the most interesting indicators for consecutive categories of job offers. By exploring the interactive visualizations of highly associated terms for selected adjectives, we have examined the considerable features of the collection of texts regarding remote work job offers.

References

Aggarwal CC., Zhai C., *A survey of text clustering algorithms*, [in:] C.C. Aggarwal, C. Zhai (eds.), *Mining Text Data*, Springer, March 2012, pp. 77-128.

Agrawal R., Batra M., *A detailed study on text mining techniques*, "International Journal of Soft Computing and Engineering" , January 2013, vol. 2, iss. 6, pp. 118-121.

Baker P.M., Moon N.W., Ward A.C., *Virtual exclusion and telework: barriers and opportunities of technocentric workplace accommodation policy,* "Work" 2006, vol. 27, no. 4, pp. 421-430.

Blansché A., Cojan J., Dufour-Lussier V., Lieber J., Molli P., Nauer E., Skaf-Molli H., Toussaint Y., *Table 3: Adaptation of Ingredient Quantities and of Textual Preparations*, 18th International Conference on Case-Based Reasoning, Alessandria, Piemonte, Italy 2010, p. 195.

Bloom N., Liang J., Roberts J., Ying Z.J., *Does Working from Home Work? Evidence from a Chinese Experiment*, CEP Discussion Paper no. 1194, March 2013, pp. 1-35.

Bolasco S., Canzonetti A., Capo F.M., della Ratta-Rinaldi F., Singh B.K., *Understanding text mining: A pragmatic approach*, [in:] S. Sirmakessis (ed.), *Knowledge Mining*, vol. 185, Studies in Fuzziness and Soft Computing, Springer 2005, pp. 31-50.

Busch E., Nash J., Bell B.S., *Remote work: An Examination of Current Trends and Emerging Issues*, Ithaca, Center for Advanced Human Resource Studies, Cornell University, 2011, pp. 9-10.

Don A., Zheleva E., Gregory M., Tarkan S., Auvil L., Clement T., Shneiderman B., Plaisant C., *Discovering interesting usage patterns in text collections: Integrating text mining with visualization*, [in:] *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Managemen*t, 2007, pp. 213-222.

Dutcher E.G., Saral K.J., *Does Team Telecommuting Affect Productivity? An Experiment*, MPRA Paper no. 41594, Webster University Geneva, September 2012, pp. 1-29.

Fuhr J.P., Pociask S., *Broadband and telecommuting: Helping the U.S. environment and the economy*, "Low Carbon Economy" 2011, vol. 2, no. 1, Scientific Research Publishing, pp. 41-47.

Garrett R.K., Danziger J.N., *Which telework? Defining and testing a taxonomy of technology-mediated work at a distance*, "Social Science Computer Review", Spring 2007, vol. 25, no. 1, pp. 27-47.

Gharehchopogh F.S., Khalifehlou Z.A., *Study on Information extraction methods from text mining and natural language processing perspective*s, "AWER Procedia Information Technology & Computer Science" 2012, vol. 1, pp. 1321-1327.

Gottfried K., *The World of Work: Global Study of Online Employees Shows One in Five (17%) Work from Elsewhere*, Ipsos 2012, press release 23.01.2012, http://www.ipsos-na.com/news-polls/pressrelease.aspx?id=5486 [accessed: April 2013].

Gupta V., Lehal G.S., *A Survey of text mining techniques and applications*, "Journal of Emerging Technologies in Web Intelligence" , August 2009, vol. 1, no. 1, pp. 60-76.

http://www.careerbuilder.com/Jobs/Keyword/Remote/ [accessed: April 2013].

http://momstowork.jobamatic.com/a/jobs/find-jobs/q-Remote [accessed: April 2013].

http://www.simplyhired.com/a/jobs/list/q-remote/fjt-telecommute [accessed: April 2013].

http://www.remoteemployment.com [accessed: April 2013].

http://jobsearch.monster.com/search/?q=remote [accessed: April 2013].

Jusoh S., Alfawareh H.M., *Techniques, Applications and Challenging Issues in Text Mining*, "IJCSI International Journal of Computer Science Issues", November 2012,v ol. 9, iss. 6, no. 2, pp. 431-436.

Kaur R., Aggarwal S., *Techniques for mining text documents*, "International Journal of Computer Applications", March 2013, vol. 66, no.18, pp. 25-29.

Lin D., Pantel P., *DIRT − Discovery of Inference Rules from Text,* [in:] *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2001, pp. 323-328.

Lister K., Harnish T., *The State of Telework in the U.S. − How Individuals, Business, and Government Benefit,* Telework Research Network 2011, pp. 21-25.

Mahesh T.R., Suresh M.B., Vinayababu M., *Text mining: advancements, challenges and future directions*, "International Journal of Reviews in Computing" 2010, vol. 3, pp. 61-65.

Mann S., Holdsworth L., *The psychological impact of teleworking: stress, emotions and health*, "New Technology, Work and Employment" 2003, vol. 13, no. 3, Blackwell Publishing, pp. 196-211.

Patel F.N., Soni N.R., *Text mining: A Brief survey*, "International Journal of Advanced Computer Research", December 2012, vol. 2, no. 4, iss. 6, pp. 243-248.

Radovanović M., Ivanović M., *Text Mining approaches and applications*, "Novi Sad Journal of Mathematics" 2008, vol. 38, no. 3, pp. 227-234.

Ramanathan V., Meyyappan T., *Survey of Text Mining*, International Conference on Technology and Business Management, March 2013, pp. 508-514.

Schadler T., Brown M., Burnes S., *US Telecommuting Forecast, 2009 To 2016, Information & Knowledge Management Professionals*, Forrester Research 2009, pp. 2-10.

Stroińska E., *Elastyczne formy zatrudnienia. Telepraca. Zarządzanie pracą zdalną*, Poltext, Warszawa 2012, pp. 115-134.

Su Z., Kogan J., Nicholas C., *Constrained clustering with k-means type algorithms*, [in:] M.W. Berry, J. Kogan, *Text Mining: Applications and Theory*, Wiley, 2010, pp. 81-103.

Teh B.H., Ong T.S., Loh Y.L., *The acceptance and effectiveness of telecommuting (work from home) in Malaysia*, Global Conference on Innovations in Management, London 2011, pp. 34-51.

Vidhya K.A., Aghila G., *Text mining process, techniques and tools: An overview*, "International Journal of Information Technology and Knowledge Management" 2010, vol. 2, no. 2, pp. 613-622.

Ward N., Shabha G., *Teleworking: An assessment of socio-psychological factors*, "Facilities" 2001, vol. 19, no. 1-2, Birmingham, pp. 61-71.

Watad M.M., Jenkins G.T., *The impact of telework on knowledge creation and management*, "Journal of Knowledge Management Practice" , December 2010, vol. 11, no. 4, pp. 237-251.

TEXT MINING W PRAKTYCE: ODKRYWANIE WZORCÓW W TEKSTACH OFERT PRACY ZDALNEJ

Streszczenie: Niniejszy artykuł ma na celu przedstawienie technik *text mining* na przykładzie nieuporządkowanych zbiorów tekstowych ofert pracy zdalnej. Przeanalizowano w nim pięć najbardziej popularnych i najczęściej odwiedzanych portali internetowych, zawierających oferty pracy, aby wydobyć przydatne informacje, odkryć ciekawe wzorce w danych oraz wyłonić cechy telepracy. Przebadano również uzyskane klastry ofert pracy zdalnej, słowa kluczowe opisujące te klastry, a także relacje między silnie powiązanymi terminami występującymi w różnych ofertach pracy. Dynamiczny rozwój narzędzi *text mining* umożliwia eksplorację nowych tematów badawczych, związanych m.in. z Internetem, freelancingiem i rynkiem telepracy.

Słowa kluczowe: *text mining*, eksploracja danych tekstowych, klasteryzacja, drzewa powiązań, praca zdalna, telepraca.