# Automatic Keyword Extraction on Amazon Reviews

Author: Ruiqi Chen
Supervisor: Jyrki Nummenmaa

# 1.Introduction

## 1.1 Background

The capacity of web content is increasing rapidly as the development of information technology. Nowadays people can create quintillion bytes of data everyday in all kinds of forms, for example, news, articles, advertisement, reviews and etc. As is known that mankind is entering into a new era of Big Data, it is not difficult to realize how data is changing our lives. In 2009 google has successfully predicted the diffusion regions of H1N1 virus a few weeks before it hit the headlines[1]. They made several correlation models based on user search queries and the results turned out to be even more timely than official announcements. Public health is not the only area where big data can make a difference, other industries like education and engineering are putting progressively focus on the research of big data too. There are quite a lot theories and conclusions have been done on big data, however, one of the most widely accepted theories is the 3V's of big data: Volume, Velocity, and Variety[2].

Big data is bringing significant convenience to people's work and life. Donald Trump posted a note on twitter and a second later millions of twitter users can know what he said. Companies like google provide people with a easiest way to search information on internet. From push service of mobile application people can always acquire worldwide news in the first time. Big data is changing the world, however, the massive amount of information that people are creating everyday make it difficult to be manually processed. The first problem is the limit of processing speed. Some real-time data required to be processed in time, otherwise it would be no longer useful. For example the stock data, a decision made on a piece of outdated stock data will

probably lead to a huge loss. In addition, human resource is no more enough for handling the growing amount of data. As reported by David Sayce, around 52 million tweets were produced every day in 2016[3]. Such volume of data is impossible for human to process. Moreover, although data amount is getting larger, valuable data only occupies a small fraction of the whole, hidden among other useless data. To solve the abovementioned problems, a robust and universal tool is strongly needed to discover the information behind the data, and transfer them into organized knowledge. This is how data mining was born. Data mining aims to quickly find the potential knowledge and possible correlations inside a data source and then apply them into every aspect of human life.

In many occasions data mining can be seen as a synonym of knowledge discovery, which generally consists of several sub-tasks: Data cleaning, Data integration, Data Selection, Data Transformation, Data mining, Pattern evaluation and Knowledge presentation[4]. The data that can be mined could be in different forms, such as numbers, texts, images or even videos. This thesis mainly focus on the keyword extraction on english review texts.

Online product review as an important portion of web data, has received a lot of research attentions recently. In the meantime, the phrase “review mining” has gradually been widely used. Review mining, which is also called opinion mining, aiming to extract critical consumer opinion towards a product from the massive unstructured review texts. In this respect, researchers such as Hu and Liu[5][6][7], Popescu and Etzioni[8][9] have made great contributions to review mining. This thesis will also focus on the analysis of product reviews, meanwhile introduce the concept of opinion ranking based on topic influence.

## 1.2 Research Question

The aim of this thesis is to develop a quick tool to analyze product reviews. First, a collection of Amazon product reviews will be crawled from internet and stored into local database. Then, some NLP(Natural Languages Processing) techniques will be applied to the data and the expect output is a list of expressions. Finally, the results will be compared and evaluated. Therefore, the research questions of this thesis can be the following:

1. How to define appropriate patterns to extract the candidate expressions from consumer reviews?
2. How to find the most important expressions and make sure they are semantically independent?

## 1.3 Research Task

This thesis mainly focus on the mining of Amazon reviews. Given all the reviews of a single product, the proposed algorithm is expected to summarize the reviews into several expressions. These expressions should consist of Nouns, adverb(optionally), and adjectives, and should be in an order based on importance. To achieve the goal, several NLP tools will be applied to the review text. Generally, statistical characteristics of the words are often used when extracting feature word and opinion word. Statistical characteristics include TF(Term Frequency), IDF(Inverse Document Frequency), word's first occurrence, word's length and etc. This kind of characteristics are easy to acquire, but have some limitations on more complex tasks. Thus, semantic information of the word is also needed to overcome the problem. Semantic information of word include POS(part-of-speech), word's synonym, dependency relations and etc. In this thesis, word's TI(Topic Influence) is used to extract the expressions. To get word's TI score, LDA(Latent Dirichlet Allocation)[10] will be employed to the review text. TI score can reflect how much a word contributes to the whole document, and it will be further used as edge weight in an undirected weighted graph. After that Textrank algorithm will be applied on the graph to rank the results. Different from traditional Textrank algorithm which takes a single word as a node, this thesis creatively takes a phrase as a node. Finally, top k expressions will be selected as expressions of the reviews. In this thesis, reviews from three different products will be analyzed in the experiment, using the proposed combination method of topic modeling and graph to extract the expressions.

## 1.4 Time Schedule

The algorithm should be developed by the end of August, 2017.
The document should be finished by the end of November, 2017.

# 2. Literature Review

## 2.1 Review Mining

To start with the discussion, some information on review mining is introduced. Similar to data mining, review mining can be divided into several sub-tasks. Popescu and Etzioni[8] define four general steps of review mining:1) Extract product features. 2) Identify opinion words related to features. 3) Calculate polarity of opinions. 4) Summarize the results. Figure 1 describes the processes in a flow chart. However, in this thesis, step 1,2,4 are mainly focused because the purpose of this thesis is to identify keywords which can summarize the reviews, the polarity

information will also be involved in the results. Additional, in the big data context, a brief literature review on data collection will also be performed.

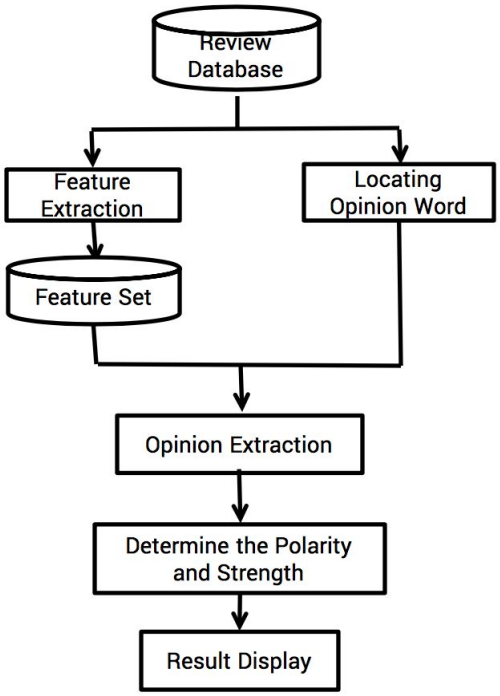


Fig 1 Framework of product review mining

## 2.2 Data Collection

As the first step of data mining, data collection is always a crucial and necessary procedure. Nowadays, there are plenty of public datasets online for researching use, such as SNAP[11], a huge amazon reviews dataset including around 35 millions reviews, and OpinRank[12], which contains cars and and hotels reviews collected from Tripadvisor and Edmunds. Such datasets can be acquired easily, eliminating the need to obtain data separately. However, most of these datasets are lack of maintenance and update, which means the data in it might be out of date. Therefore, more researchers choose to get their own data to ensure the data quality. In general, the collecting of the data is done by a system called web crawler. For simple single-format data, the crawler can be very lightweight. Hu and Liu[5], Kasper and Vela[13], Owsley and Sood[14] collect the reviews by a customized web crawler and then they store the data in a local database. For large scale multi-format data, a more comprehensive and complex crawler has to be developed. In this respect, Myllymaki[15] developed an XML based system ANDES, which can crawl relevant websites through a seed website and then extract domain-specific content from massive html structures. Chau and Pandit[16] proposed a parallel mining model, in their model a central computer controls the mining task queue and assign the tasks to different agents, the

agent then execute the task multithreaded. They tested their model on an online auction website and greatly reduced the processing time. Similarly, Cheng[17] divide the large scale data mining task into several small tasks, and then run them paralleled on different server in order to improve efficiency.

Since the data of this thesis is only a small amount of amazon product reviews, a customized web crawler is enough to accomplish the task.

## 2.3 Feature Extraction

In most ecommerce websites, the product page often contains a short product description from the manufacture. However, this kind of description is not a suitable resource for review mining, although it may involve information about product features. The reason is that, manufactures may have different concerns about product feature with the consumers. Some electronic manufactures like to provide information on technical details, for example, a mobile phone manufacture will probably focus on describing the performance of the kernel, while most of the consumers are more concerned about the running speed when having a lot of applications installed. In addition, the manufacturer's description of the product is not comprehensive, some product features mentioned in user reviews are not taken into account by the manufacture. Thus extracting the features from reviews is very necessary.

Product feature extraction is a crucial process of review mining. It aims to extract the product aspects which the consumers have made expressions on. Features are usually in the forms of noun or noun phrase. Yi and Niblack[18] believe a feature must meet one of the following three conditions: 1) It has to be part of the giving subject. 2) It has to be an attribute of the giving subject. 3) it has to be an attribute of a part of the giving subject. Taking mobile phone as an example, the screen is a product feature, since it is a part of the phone;  The price is also a feature because it is an attribute of the phone; The image quality is a feature, because it is an attribute of phone camera, which is a part of the phone.

Product features are generally divided into explicit features and implicit features[5]. As their name suggest, explicit features refer to the features that show explicitly in a sentence, while implicit features refers to the features that do not show directly in a sentence, but need a deep-level understanding of the sentence to know. Following two review sentences are extracted from Amazon website as an illustration:

"I LOVE this camera - easy operation, great pictures. fantastic price. "

"It's small enough to throw in my purse and easy to use."

In the first sentence, it is easy to know that words "operation", "pictures" and "price" are explicit features. In the second sentence, there is no such noun or noun phrase could be taken as a feature, only after understanding the the whole sentence it can be inferred that the author is talking about the size of the camera. It can be seen that extracting Implicit features are more challenging than extract explicit features. This thesis mainly focus on explicit feature extraction, leaving the implicit feature extraction to future development.

There are two ways for extracting explicit features, which are manual definition and automatic extraction. Manual definition is to set up a feature vocabulary for products from a specific area. In the respect, Zhuang et al[19] define several classes(screenplay, character design, vision effects, actor and actress, etc) for movie features by observing the reviews from IMDB, and then use a statistical method to determine the movie feature set. Blair et al[20] use a combinational approach of manual definition and automatic extraction to extract the features from local service reviews. They define four features(food, decor, service, value) for restaurant and five features(rooms, location, dining, service, value) for hotel. For each of the feature set they merge them together with auto-extracted features to improve the overall accuracy. Yao et al[21] develop a supervised system for automobile based on a manually created ontology base, their system comprehensively analyzed the opinions towards different features of one car as well as one feature from different cars. Kobayashi et al[22] also develop a semi-automatic system for collecting opinion expressions from game and automobile reviews. Given three manually selected seed sets of subjects(products), attributes(features) and values(opinions), their system will extract the evaluative expressions based on predefined cooccurrence patterns. However, a human judger is still needed to evaluate the expressions in the final step.

However, there are some drawbacks to manual definition of product feature. Firstly, with the rapid growth of the world economy, the variety of products is also increasing quickly, which means manual definition becomes especially unrealistic to cover all the domains. Secondly, the manufactures often need to update their product design according to market research, while the manually defined features still remain unupdated, leading to an inaccurate results of experiment. Meanwhile, different domain experts are needed to create domain-specific features, which is a huge cost of time and money.

Automatic product feature extraction mainly employs the natural language processing techniques such as part-of-speech tagging, syntactic analysis, document pattern and etc. Given a sentence, automatic feature extraction can locate the feature word based on some restrictions and predefined rules. Both supervised approach and unsupervised approach can be used to accomplish the task.

For supervised learning, Hu and liu[6] manually label feature words that occur in the reviews, for convenient they separate the sentence into 3-gram segments and save the segments in a transaction file. They then apply association rule mining on the file to acquire common patterns, which can be used to identify possible features in new reviews. Kessler et al[23] focus on finding semantic relation between feature words and opinion words. They annotate both features and opinions in a dataset of car and digital camera reviews. Supervised machine learning are employed to rank possible features linked to an opinion word. Their algorithm yields a precision of 0.748 and recall of 0.654, both are higher than the baseline algorithm, which is proposed by Bloom et al[24] in 2007. Supervised approaches normally perform well on review mining, yet one disadvantage is the need for manual labeling in advance.

For unsupervised learning, Hu and Liu[5] apply POS tagging on review sentences and save the noun/noun phrase in a transaction file. An association miner is again used on the file to extract frequent features. Compactness pruning and redundancy pruning are used to filter the result. The system can also identify infrequent features by checking if opinion words exist in the same sentence. Their system can extract the features from multi-domain reviews, however, the recall and precision of the algorithm are only 0.693 and 0.642.

Kim and Hovy[25] employ a semantic role labeling approach to extract the topic(feature) and opinion holder from a sentence. First opinion word is extracted from the sentence and a frame class is assigned to the sentence based on FrameNet data. They then label the sentence fragments with their semantic roles using a statistical method. A mapping between the semantic roles with opinion holder and topic(feature) is created manually to identify the feature and holder of the given opinion word. Their system yields an average precision of 0.618 on topic(feature) extraction, which is much higher than the baseline, which yields only 0.179. However, their system depends a lot on external corpus, causing a risk of unstableness in future development.

On top of Know-it-all system, Popescu et al[8] develop an unsupervised review mining system called OPINE. Given an input of product class and pre-defined rule templates, the system will extract candidate features based on the rules. To improve the extraction accuracy,

PMI(Point-wise Mutual Information) score, which depends on the hit counts from web searching, is calculated for each of the candidate in order to check the probability of it being a feature of the given product class. Their system receives a 22% higher precision over Hu and Liu's algorithm on the same dataset, while only has a 3% lower recall. However, since calculating PMI will consume a lot of time, their system is not suitable for large dataset mining.

这个section和我研究的内容高度相关,需要多看细看这里

## 2.4 Opinion Extraction

Opinion word refers to the word which the author uses to express her/his feeling about a certain product feature. Some researchers extract opinion words using an opinion words dictionary. For example, Zhuang et al[19] select top 100 positive words and negative words with highest frequency from their labeled training data, and take these opinion words as seed set. In order to find unobserved opinion words in training data, they iterate through WordNet and find the words with as least one seed word exists in their synsets, and then add these words into final opinion words list. Finally they are able to extract the opinion words based on the opinion words list. Lun at el[26] try to create a chinese opinion words dictionary for news and blogs. They first collect the opinion words from GI(General Inquirer) and CNSD(Chinese Network Sentiment Dictionary) and then take these words as seed set. They then expand the opinion words by searching for their synonyms in Cilin(TongYiCiLin) and BOW(Academia Sinica Bilingual Ontological Wordnet). In addition they calculate polarity score for each word based on a positive formula and a negative formula.

Another approach to opinion word extraction is to discover the relations between feature words and opinion words. By observing the reviews, Hu and Liu[5] find that opinion word usually occur near to the feature word. According to this observation they collect the opinion words by checking if adjectives exist nearby the feature word. For example, in the review "The appearance of this phone is good.", they first locate the feature word "appearance" and then find the nearest adjective "good". This approach is easy to implement, however, it only considers adjective as opinion word, ignoring that some verbs and adverbs may also express the author's attitude. For example, in the review "I love this phone.", the word "love" indicates the semantic orientation too. Inspired by Hu and Liu's work, Popsecu et al[8] manually define 10 dependency relations between feature word and opinion word based on the parsed result from MINIPAR parser. Their algorithm can not only detect adjective opinion words, but also noun and verb opinion words. However, opinion words that do not meet the rules will not be detect from the reviews. In another paper proposed by Hu and Liu[7], they focus on analyzing the reviews in the form of "pros" and "cons". Such kind of reviews are commonly exist in Amazon website. They develop a supervised method to mine the CSRs(Class Sequence Rules) from labeled reviews. The rules can then be used to identify feature words and opinion words in a review sentence. Feng et al[27] extract the feature-opinion pairs based on some dependency rules. They first parse the review

text using Stanford Dependency Parser, and then extract the word pairs with three main dependency relations, including “amod”, “nsubj” and “dobj”. Likewise, their algorithm can also detect verb opinion words. Yi et al[28] design a system for review mining, which called SA(Sentiment Analyzer). SA first extract feature words from review sentence, and then extract the ternary expressions in the form of <target, verb, source> as well as binary expressions in the form of <adjective, target>. By using several external sentiment lexicons the system can calculate the polarity of the each expression. Dini et al

## 2.5 Scoring and Ranking

After acquiring the expressions that contain feature word and opinion word, it is necessary to sort them by importance. An expression with higher score usually means more concerns are put on the corresponding product feature. By ranking the expressions consumers can easily understand the parameters of the product, checking if it meets their expectation. In addition, product designers may have strong interests on the sorted result, which can tell them what most consumers care about. Meanwhile, by filtering out the expressions with lower score, the accuracy of the result can be improved.

A lot of researches have been focusing on keyword extraction and ranking. A very fundamental way to get keywords is to count the number of occurrence of each unique word in the document, and taking top k words with highest frequency as keywords. Based on this concept, one of the most famous approach is TFIDF, proposed by Salton and Buckley[29] in 1988. TF can reflect the capacity of a certain word describe the documents, while IDF can reflect the capacity of a certain word distinguish the documents. The concept of TFIDF is that, when a word occurs many times in one document but seldom occurs in other documents means this word has a strong capacity to represent the current document, and a word with higher TFIDF score will be more important. The drawback of TFIDF is also obvious, since it only use the statistical information of word, ignoring the semantic information behind the document.

Rose et al[30] develop a rapid automatic keyword extraction method for individual document. They calculate the word weight based on the word degree as well as the word frequency. For multiple word expressions they calculate the weight by summing the members’ weights up. Their approach proved to be very efficient and universal.

Recently, Lda-based keyword extraction gradually received people’s favor[31][32][33]. Inspired by human writing styles, lda believes a document is a mixture of several topics, while a topic consists of a list of words. Lda model can be trained by a document corpus. The output of lda algorithm is a document-topic distribution and a topic-word distribution. By calculating these

two distributions a document-word distribution can be inferred, which reveals TI score of each word. A ranking based on TI score can then be easily made.

Furthermore, graph-based keyword extraction has also yield great success[34][35][36]. The basic concept is to regard the document as a word based network. TextRank[34] is one of the most famous algorithm in this area. Inspired by PageRank, TextRank takes words as the nodes of the graph, by setting a fix sized window the algorithm checks if two words co-occurred in the window. If yes, add an edge between these two words in the graph. The algorithm will output the score of each node in the graph.

On the basis of previous chapters, this thesis decides to adopt a method similar to Feng's approach to extract the expressions from review sentences. Besides, the thesis will calculate the TI score of each expression, and take it as edge weight to create a undirected weighted graph. Finally, running TextRank algorithm on the graph to get final sorted result.

## References


[1]. Mayer-Schönberger, V. (2013). "Big data : a revolution that will transform how we live, work, and think".

[2]. VK Jain, V. (2017). "Big Data and Hadoop".

[3]. David Sayce. [online] David Sayce. Available at: https://www.dsayce.com/social-media/tweets-day/ [Accessed 20 Sep. 2017].

[4]. Han, J., Pei, J. and Kamber, M. (2012). Data Mining: Concept and Techniques. 3rd ed. Pp.8,9,10.

[5]. Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp.168-177.

[6]. Liu, B., Hu, M., & Cheng, J. (2005). Opinion Observer: Analyzing and Comparing Opinions on the Web. *Proceedings of the 14th international conference on World Wide Web - WWW 05.*

[7]. Hu, M., & Liu, B. (2006). Opinion feature extraction using class sequential rules. *Computational Approaches to Analyzing Weblogs, Papers from the 2006 AAAI Spring Symposium, Technical Report SS-06-03, Stanford, California, USA.*

[8]. Popescu, A. and Etzioni, O. (2005). Extracting Product Features and Opinions from Reviews. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing.*

[9]. O. Etzioni, M. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D. Weld, and A. Yates. (2005). Unsupervised named-entity extraction from the web: An experimental study. *Artificial Intelligence, 165(1):91–134.*

[10]. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research.*

[11]. McAuley, J. and Leskovec, J. (2013). Hidden factors and hidden topics: understanding rating dimensions with review text. *Proceedings of the 7th ACM conference on Recommender systems, pp.165-172.*

[12]. Ganesan, K. and Zhai, C. (2011). Opinion-Based Entity Ranking. *Information Retrieval,* 2011.

[13]. Kasper, W. and Vela, M. (2011). Sentiment Analysis for Hotel Reviews. *Proceedings of the Computational Linguistics-Applications Conference, 2011, pp.45-52.*

[14]. Owsley, S., Sood, S. and Hammond, K. (2006). Domain Specific Affective Classification of Documents.

[15]. Myllymaki, J. (2001). Effective Web data extraction with standard XML technologies. *Proceedings of the 10th international conference on World Wide Web, pp.689-696.*

[16]. Chau, D., Pandit, S., Wang, S. and Faloutsos, C. (2007). Parallel Crawling for Online Social Networks. *Proceedings of the 16th international conference on World Wide Web, pp.1283-1284.*

[17]. Cheng, M. (2011). Web Data Mining Based on Cloud-computing. *Computer Science, 2011.*

[18]. Yi, J. and Niblack, W. (2005). Sentiment mining in WebFountain. *Proceedings of the 21st International Conference on Data Engineering, pp.1073-1083*.

[19]. Zhuang, L., Jing, F. and Zhu, X. (2006). Movie Review Mining and Summarization. *Proceedings of the 15th ACM international conference on Information and knowledge management, pp.43-50*.

[20]. Blair-goldensohn, S. and Hannan, K. etc. (2008). Building a sentiment summarizer for local service reviews. *Proceedings of the WWW2008 Workshop: NLP in the Information Explosion Era (NLPIX 2008)*.

[21]. Yao, T. and Nie, Q. (2006). An Opinion Mining System for Chinese Automobile Reviews.

[22]. Kobayashi, N., Inui, K., Matsumoto, Y., Tateishi, K., & Fukushima, T. (2005). Collecting Evaluative Expressions for Opinion Extraction. *Natural Language Processing – IJCNLP 2004 Lecture Notes in Computer Science, pp. 596-605.*

[23]. Kessler, J. S., & Nicolov, N. (2009). Targeting Sentiment Expressions through Supervised Ranking of Linguistic Configurations. *Proceedings of the Third International ICWSM Conference.*

[24]. Bloom, K., Garg, N., & Argamon, S. (2007). Extracting Appraisal Expressions.

[25]. Kim, S., & Hovy, E. (2006). Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. *Proceedings of the Workshop on Sentiment and Subjectivity in Text , pp. 1-8.*

[26]. Ku, L., Liang, Y., & Chen, H. (2006). Opinion Extraction, Summarization and Tracking in News and Blog Corpora.

[27]. Feng, S., Zhang, M., Zhang, Y., & Deng, Z. (2010). Recommended or Not Recommended? Review Classification through Opinion Extraction. Advances in Web Technologies and Applications, *Proceedings of the 12th Asia-Pacific Web Conference, APWeb 2010, Busan, Korea, 6-8*.

[28]. Yi, J., Nasukawa, T., Bunescu, R., & Niblack, W. (2003). Sentiment Analyzer: Extracting sentiments about a given topic using natural language processing techniques. *The Third IEEE International Conference on Data Mining*.

[29]. Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management: an International Journal, 24(5), 513-523*.

[30]. Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic Keyword Extraction from Individual Documents. *In book: Text Mining: Applications and Theory, 1-20*.

[31]. Ma, B., Zhang, D., Yan, Z., & Kim, T. (2013). An LDA and Synonym Lexicon Based Approach to Product Feature Extraction from Online Consumer Product Reviews. *Journal of Electronic Commerce Research; Long Beach Vol. 14, Iss. 4, (2013): 304-314*.

[32]. Shi J. and Li W. (2010). Topic Words Extraction Method Based on LDA Model. *Computer Engineering, 36(19): 81-83.*

[33]. Liu J., Zou D. & Xing X. (2012). Keyphrase Extraction Based on Topic Feature. *Application Research of Computers, 29(11): 4224-4227.*

[34]. Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing Order into Text.

[35]. Litvak, M., & Last, M. (2008). Graph-based keyword extraction for single-document summarization. *Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization,17-24*.

[36]. Tsatsaronis, G., Varlamis, I., & Nørvåg, K. (2010). SemanticRank: ranking keywords and sentences using semantic graphs. *Proceedings of the 23rd International Conference on Computational Linguistics , 1074-1082*.