ELEKTRONİK EĞİTİM BİLİMLERİ DERGİSİ
ELECTRONIC JOURNAL OF EDUCATION SCIENCES

Yıl/Year:2017
Cilt/Volume: 6
Sayı/Issue:12
ISSN:1302-7905

USING THE TECHNIQUES OF DATA MINING AND TEXT MINING IN EDUCATIONAL RESEARCH

EĞİTİM ARAŞTIRMALARINDA DATA MINING VE TEXT MINING TEKNİKLERİNDEN YARARLANMA

Mustafa Ergün[1]

Öz

Günümüzde resmi ve özel verileri depolamak ve geri getirmek için birçok olanakların doğmasıyla herkes fotoğraflarını, filmlerini, yazılarını, her türlü notlarını elektronik ortamlarda saklamaktadır. Buna bağlı olarak bu ortamlarda saklanan bilgiler her geçen gün akıl almaz bir şekilde artmaktadır. Eğitim araştırmaları giderek metin, ses, görüntü temelli verilerle yapılmaya başlanmıştır. Metin, ses ve görüntü analiz etme ve inceleme tekniklerindeki gelişmeler ile metin madenciliği giderek daha çok alanda kullanılmaktadır. Metin madenciliği gelecekte daha kolay olabilir, çünkü Anlamsal web denilen Web 3.0 teknolojileri artık internetteki yazı ve nesneleri anlamlarıyla birlikte işleyip değerlendirecektir. Böylece birçok metin madenciliği işlemi daha dosya internete konulurken dosya üzerinde yapılmış olacak; metadata kelime grupları (ontolojiler), kelime ve kavram haritaları (mind maps) vs. dokümanlar içinde hazır olacak. Genel internet ağının altındaki Anlamsal Ağ'da temel amaç veriyi, iyi tanımlanmış ve ilişkilendirilmiş olarak bilgi haline getirecek ve servislerin genel ağ ortamında kolay bir şekilde makineler tarafından okunabilir ve anlaşılabilir olmasını sağlayacak standartların ve teknolojilerin geliştirilmesidir. Web 3.0 teknolojisi birçok şeyi bilen bir "asistan" rolüne girmektedir. Halen bir arama motoru üzerinden arama yapıldığında aranılan kelimeyi içeren genel ağ siteleri sıralanmaktadır. Anlamsal Ağ etkin olarak kullanıldığında ise, makineler web sayfalarını yorumlayabilecekleri için doğru sonuçlara ulaşılmasını sağlayacaktır. Kaynak Tanımlama Çerçevesi (Resource DF) bir kelimeler hiyerarşisi ve nesnelerin ilişkilerini dosya içine yerleştirecektir.

Anahtar kelimeler: eğitim araştırmaları, veri madenciliği, kodlama, metin madenciliği, kavram haritası, güvenilirlik.

Abstract

Today there are many specialized ways of data storing and retrieving. Photographs, movies, articles and other similar data can be stored in electronic settings. Given that information stored in such settings is increasing incredibly. In educational research the use of text-, audio- and visual-based data has become frequent in recent years. In parallel to this change new techniques have appeared regarding such data. It is thought that text mining will be easier in future, because the Web 3.0 technologies, also called semantic web, would analyse texts and objects in internet together with their meaning. In other words, the texts would be put into Internet after the use of text mining. Metadata vocabulary groups (ontology), groups of words and concepts (maps) will be ready in texts. The basic aim of the Web 3.0 technology or semantic web is to make data usable and connected information. The Web 3.0 technology would assume the role of assistant that knows everything. At present when we search on the web, general websites that contain the related word or words are listed. When the Web 3.0 technology or semantic web is used effectively, the engines will interpret the websites and only related websites will be offered. The Web ontology language will be standard, and it identifies common vocabulary or terminology. The resource description framework (RDF) will be put in files containing vocabulary hierarchy and the correlations of the objects.

Keywords: educational research, data mining, text mining, coding, concepts map, reliability.

[1] Prof. Dr., Afyon Kocatepe Üniversitesi Eğitim Fakültesi, ergun@aku.edu.tr

1. INTRODUCTION

Today it is possible to store and retrieve official and personal data; everyone can store photographs, movies and notes in electronic settings. The amount of such data increases everyday (some call it "information boom"). Many search engines can scan electronic data based on key terms (or based on content); some sales companies also use such data to offer "best" sales. Some computer programs provide their clients with textual data collected from Internet sources. Such programs are mostly used by sales companies. Today emails, notes on facebook and twitter as well as blogs are checked for security purposes. In science information-based search engines such as GoPubMed are employed. In short, computer-assisted text analysis has been used in different fields and various techniques make it possible to analyse data.

Most of the data stored at libraries and digital settings are unstructured texts, voice and picture files. Using such data as information for researchers is very significant in terms of research methods. Techniques and software are needed to analyse these data correctly and systematically.

Educational research is increasingly based on the textual, audio-visual data. In parallel to this development, the needed research techniques have been developed. On the other hand, audio and textual data should be simultaneously analysed, but films or photographs should be separately taken into consideration.

1.1. Content analysis

Content analysis is one of the traditional research methods used in social sciences and media studies. It may use both qualitative and quantitative data. It has many subcategories, including sentence analysis, speech or discourse analysis, thematic analysis, and vocabulary analysis. In these techniques the points taken into consideration include the frequency of words, word associations, conceptual associations, differential use of concepts by individuals, and grouping of words. Following the coding the process called thematic analysis takes place. There are three major types of content analysis. The first one is the traditional content analysis in which coding categories are extracted from the text. The second one, called directed content analysis, tries to find out codes in accordance with theory or hypotheses and analyses these codes. The third one is summative content analysis which looks for, compares and interprets key words (Hsieh & Shannon, 2005).

Both descriptive and interpretive research designs can be employed. Descriptive analysis is mostly used when the data are strictly used. If the data are to be freely interpreted, interpretive designs are employed. Content analysis is much more frequently used in descriptive designs, while thematic analysis is much more common in interpretive designs. Content analysis identifies the codes and thematic analysis produces concepts and themes based on these codes and then, interprets them.

Content analysis is carried out through Excel-based software for the data obtained from structured interviews and survey questionnaires. Here frequency and percentages are identified, and cross tabulation is developed as well as chi-square is calculated to make sound interpretations. The data can be presented through graphics.

Both people and computer can better process and use structured and/or listed information. The examples of the structured data include student records, hospital records, bank records, etc. In such information the related data about a person or an event is recorded and listed. Information stored in data bases developed through SQL or Access can easily be searched, found, and classified (Akpınar, 2000). It is discussed whether or not the information

stored in the Excel programs are structured. Such data can be searched in different ways to get useful information.

Computers generally produce "related terms" based on the morphological characteristics of words rather than the meaning of words. The textual movement of the relations obtained using this way can be presented through graphics.

Some of the software used for content analysis include Concordance, Crawdad Technologies LLC, Diction, General Inquirer, Hamlet II, INTEXT, Leximancer, Minnesota Contextual Content Analysis (MCCA), Profiler Plus, PROTAN, SALT Software, PASW Text Analytics for Surveys, T-LAB Tools for Text Analysis, TABARI, TACT (Text Analysis Computing Tools), Tapor Tools, Text Analysis Tool, Textalyser, Textanz, TextArc, TEXTPACK, TextQuest, VBPro, Wordcruncher, WORDij, Wordle, WordStat, Yoshikoder.

In this field only, limited number of texts could be used. However, it is needed to examine the different types of texts and audio-based and visual information.

1.2. Data Mining

Search engines give us many related or unrelated data. However, it is hard to access the related ones. Given that such data are not structured, it takes longer time to access the related data. This information is of text written in natural sound, visuals or language. It is estimated that more than 80% of the computer-based data are unstructured. Numerous firms, and institutions has stored their records at electronic settings and become "rich in terms of raw data, but poor in terms of having information". For these firms and institutions it is hard to be successful in a competitive market and to maintain their success (Dolgun et. al. 2009).

Natural language processing (NLP) is particularly hard in terms of semiotics. It is not enough to group words. Instead, the prefixes and suffixes attached to the words should be identified as well as the the meaning of the words in certain contexts should be determined. All these procedures require the use of much more elaborated techniques. For instance, life other words the word "cold" has different semantic relations (it may use to modify the words including water, air, behaviour). Therefore, the NLP employs artificial intelligence and linguistics as tolls to analyse and understand texts (Shi et. al. 2014). There is computer software which can be used in natural language processing.

The data stored in computers should be processed to use in the analyses. The process begins with the extraction of certain words and names to make the text more structural. If it is known what we look for in the text and how to evaluate the data, it is possible to employ the Information Extraction (IE). Therefore, the IE refers to reveal the textual meaning and to make it understandable by computer (Arslan 2011). The IE uses the extraction of names and related items (the names of persons, institutions, places, time periods, money amount and abbreviations used in the texts) and the relational extraction (connections and disconnections).

Data mining extracts the information in the text, but it is needed first to make the text structural (it is mostly done through text mining). Data mining used for the structural texts has several techniques as follows (Şentürk, 2006, Savaş et. al. 2012):

1. Classification: the data are decomposed and placed into certain categories. Therefore, the relationships among both dependent variables and independent variables are examined; at the same time new information can be discovered or constructed.

2. Clustering: The data coded are classified in the form of groups and classes and are decomposed. There are different clustering algorithms that can be used for different data types. Researchers determine which variables would be used as the basis for clustering.

3. Association rules and sequential patterns: These rules and patterns reveal the relationships among the data sets. The elements in the data which are frequently repeated forms the association rules. These rules are the most needed information. It can be achieved through the data analysis programs based on artificial intelligence. Such information can be visualized via certain software.

In data mining people and computers/software work together. People scan and code the data and produce, question and learn the rule sets to the computer (computer-learning approach). Then, computers automatically evaluate and classify each newly added data based on the rules and patterns.

1.3. Text Mining

Although text mining appeared in the 1980s, it has improved through technological advances. It is also called "text data mining" or "text analytics" as well as "concept mining" and "web mining".

Web mining converts unstructured web contents into structured ones and analyses the data about websites such as page patterns and web statistics. In such activities generally web pattern mining, web content mining and web use mining are employed.

Text mining is a subpart of the data mining, which deals with the unstructured data, which are stored in digital settings in the forms of language, sound and visuals and which are available for processing. There is an interactive relationship between text mining and data mining. The structured data obtained from the text mining are evaluated using the models of the data mining and the findings are used to analyse the textual structure.

Text mining is a method used to reveal the meaning of the textual data. Given that there is no standard rules in producing texts, computers cannot directly analyse these texts. Each text has its own language, meaning and purpose.

Therefore, text mining deals with the classification of texts as a whole, summarising the texts, revealing the representative words and concepts, comparison of similar texts and revealing the relationships among texts (Çalış et. al. 2013). However, it is not so easy; because meaning of words and hence, texts can indefinitely vary, leading to misconceptions of it (Cohen & Hunter, 2008).

Analysing unstructured data can also be carried out using traditions methods, including key terms or local connections (looking for connectives such as and, or), statistical or probability-based algorithms, discovering patterns, which are all non-linguistic methods (Dolgun et. al. 2009). In other words, first key words and concepts are found in the text, then these are grouped under higher categories (i.e., grouping all words related to computer hardware under the category of "computer hardware"). On the other hand, based on key terms the text can be summarized (for instance, grouping the letters of complaint based on key words). Then, the latent patterns can be discovered and models that can make estimations about the course of events or the acts of individuals. In fact, one of the goals of data mining is to develop predictions based on the connections and rules discovered.

In text mining the labelling and classification are done suing certain algorithms, which are taught to computers. The most frequently used algorithms include Naive Bayes (NB), k-Nearest Neighbors (kNN) and Support Vector Machine (SVM). However, before using these algorithms the vector space of the groups in which textual material would be placed should be identified.

Naive Bayes decomposes emails to uncover spam mails which include the word advertisement or similar expressions. The algorithm uses estimation and produces classifications based on pre-determined words to group the files. All possibilities are evaluated based on the vectors.

The k Nearest Neighbor (kNN) algorithm is used for classification and regression in pattern recognition. In both cases, the input consists of the k closest training examples in the feature space. The computer takes the closest three models and classifies the text into the related model. When there are more than two groups taught to the computer the Support Vector Machine is used to reveal the correct group for the text.

In using these algorithms there are several points to be taken into consideration: first the text to be analysed with text mining should be carefully chosen and non-textual elements, labels and advertisements should be excluded from the text. In recent years texts in Internet have been labelled as being a text or visual. It indicates the fact that these texts were systematically recorded, but these labels may be modified to analyse the text. The frequency of words allows for evaluation, but the frequency of "stop words" (such as at, the, which) should not be taken into consideration. There may be lemmatization in the text. Although the co-occurrence–based methods were common once a time, these methods are not no more used in the analyses. Furthermore, these methods have not been seen part of text mining techniques. Now, the most frequently used techniques in the rule-based or knowledge-based approaches, statistical or machine-learning approaches. The rule-based or knowledge-based approaches search for information clusters and relations and attempt to produce information based on word coding patterns. The statistical approaches try to classify texts and sentences based on labels. There are conflicts and uncertainty among the relationships of categories due to the lexical meaning and symbols. Researcher should reveal information using the patterns in the texts or sentences.

Textual mining has been increasingly employed in many fields. For instance, in order to manage customer relations, the related information is uncovered and relations are based on this information. Patterns in textual data collected by insurance firms and government institutions are search for abnormality to identify the fraud possibility. Quality predictions are developed based on the analysis of the patient reports, economic reports, published research results and other publications. Certain concepts and models are used to detect terrorism cases, aggravated theft and criminal offenses.

Pattern discovery in text mining is like factor analysis. Like thoughtful analysis carried out to identify the factors of a scale, the textual patterns should be thoughtful analysed by a computer. Like semi-processed data on which frequency, relations are sought in quantitative, here the word frequencies and connections are determined. One of the goals is to obtain visualized forms. In textual analyses several psychological characteristics such as emotions, views can also be determined (such studies are also done using sound and photograph files).

Software is employed for interpretation, management of texts and concept extraction as well as theory development. This software (for instance, Atlas.ti) could easily handle written texts, sound files and visuals, graphical data. There is also software dealing with the discourse analysis of patients in the psychotherapy training. There are others such as the

Ethnography which analyse interviews, focus group interviews, field notes, diaries, and meeting minutes. In short, there is numerous software used in text mining as follows: Angoss, Attensity, AUTINDEX, Autonomy, Averbis, Basis Technology (which can perform analysis in more than twenty languages), Clarabridge, Complete Discovery Source, Endeca Technologies, Expert System S.p.A., FICO Score, General Sentiment, IBM SPSS Text Analytics, Intellexer, Inxight, LanguageWare, Language Computer Corporation, Lexalytics, LexisNexis, Luminoso, Mathematica, Medallia, Megaputer Intelligence, NetOwl, RapidMiner, SAS Text Miner and Teragram, Semantria, Smartlogic – Semaphore, STATISTICA Text Miner, Sysomos, Textalytics, WordStat, Xpresso… However, most of them are not available for analysis in Turkish language.

1.4. Using text mining in the analysis of scientific studies

The programs used by scientists to produce indices are part of "text mining". These programs were first used in biology and medicine. They are now used in social sciences. The statistical analyses of written materials such as books and articles are called bibliometrics. Scientometrics is part of it. It deals with numerical analysis of academic literature, including citation and content analysis, measurement of impact factor of scientific journals. These are easily and commonly done by automatic algorithms. The impact of the scientific articles are measured by the number of citations and there is a citation index called "Web of Science" by ISI. These tools measure the popularity of the authors, articles and journals as well as impact factor of each journal. Bibliometrics can also measure such characteristics as the frequency of the use of terms and the grammatical and syntactic patterns in texts. In addition to books and scientific articles Altmetrics measures the number of downloads of articles, books, videos, of covering in social and written media outlets. Informetrics measures the production, distribution and use of news. Webometrics measures the popularity of web sites.

In bibliometrics the frequent techniques include science mapping, which analyses conceptual, mental and social structures of research fields (Cobo et. al. 2011). It reflects the structure of scientific research. The co-word analysis is used to uncover the conceptual structure of documents based on major key terms, while the co-author analysis shows the connection between the author and its study, social structures and connection networks. The co-occurrence of the words is used to develop matrices and information is produced based on the position and relations of words in the texts. The co-citation analysis deals with the studies employed the same sources. Science mapping makes it possible to produce relation maps in the groupings of scientific research.

The software that allows for science mapping are as follows: Pajek, UCINET, Cytoscape, Bibexcel, CiteSpace II, CoPalRed, IN-SPIRE, Leydesdorff's Software, Network Workbench Tool, Science of Science (Sci2) Tool, VantagePoint, VOSViewer

The temporal analyses are used to discover the patterns, trends, seasonality and extremes. At the end of each period the characteristics of which frequency is higher are determined using the burst detection technique. Geospatial analysis deals with what's happening and what the effects of it are. Visualization techniques (heliocentric maps, geometrical models, thematic networks) are determined based on the similarities and closeness among items.

Scientific documents can also be analysed using the co-word analysis techniques (He, 1999). The impact of the connections among major concepts is measured to reveal the conceptual patterns. The co-word analysis deals with the co-citation of the studies cited to uncover the pattern mapping of the field. In short, the common concept sets are sought in

documents. Instead of interrelations of citations, those of key words are analysed. Therefore, the conceptual maps characterizing a discipline is revealed. To this end the relations and an algorithm were developed. The goal is to reveal the related basic fields and the impact of the correlations among terms.

Given that in text mining both conceptual extraditions and connections among concepts are very significant, such studies are called concept mining. It attempts to identify concepts based on the words in artefacts. On the other hand, the meanings of the words should not have confused each other. It tries to develop hypernymy and metonymy. Concepts are not only based on words, but also on the sentential meaning (Puri, 2011). Different controlled learning paradigms (classification of new documents based on predetermined and computerized categories) and different classification algorithms make it possible to extract the concepts in documents and to fast gather millions of data under different categories.

Upto now the conceptual mining studies mostly have used the WordNet, which is a lexical data base (Aydın et. al. 2013). In the WordNet there is a synset, which refers to the correlation sets defining the correlations of words. Such correlations include hyponyms, synonyms, and antonyms. Of them the most used one in conceptual mining is hypernymy, which refers to the general meaning of a word. In some studies, another data set, ConceptNet, is used. It looks for physical, social, temporal correlations among words.

In a study on the Turkish language concerning conceptual mining the clustering algorithm was employed. In the study the pre-determined words were grouped, and these groups were searched in the test documents to identify the concepts. The success of the method used in the study was reported to be 51% (Aydın et. al. 2013).

The Latent Dirichlet Allocation, one of the artificial intelligence algorithms, can also be used in extracting concepts from documents. It looks for the other words, which may have collocations with key words, to reveal the latent semantic patterns of sentences or text itself (Blei et. al. 2003).

1.5. Visualization

The human brain processes visual information better than it processes text -- so using charts, graphs, and design elements, data visualization can help you explain trends and stats much more easily.

Whereas data in text form can be really confusing (not to mention bland), data represented in a visual format helps people extract meaning from that data much more quickly and easily. You can expose patterns, trends, and correlations that may otherwise go undetected.

The visualization of texts refers to graphics, schemas, flow diagram and a variety of maps and atlases. Using both quantitative and qualitative data the data visualization studies employed various technologies (Friendly, 2006). Text mining carries out visualization using the tokenizing, stemming, n-gram generation, and word clouds. There are cloud studies in visualization of the data based on the word frequency, and also those studies based on the vectoral three-dimension spatial basis.

Here the significant point is to reveal the frequency of the defining words and how to represent them. It is a technique to show the frequency of words in a text and to produce an alphabetical index. The order can be based on the frequency instead of alphabetical index. The words can be coloured or given in big characters according to their frequency. On the

other hand, in 2002 J. Flanagan began to use a cloud software which shows words in graphics based on their frequency.

The word cloud or tag cloud, which was formerly called text cloud, puts words with varying sizes and colours based on the frequencies. It enables to understand the theme of the texts with a visual. Labelling cloud makes the labels visuals based on the meaning, frequency and other characteristics as well as their correlations with others using statistical techniques and mathematical models. It is possible to have a cloud of any book. Clouding also allows for noticing the differences of discourses of politicians over time. It can also make comparisons between texts. Therefore, there are various visualization techniques using multi-dimensional ways (Bilgin & Çamurcu, 2008).

The data clouds can be employed to process and develop visual analyses of the demographical characteristics, exchange data and student records. Visualization is required to analyse and make it understandable the huge data sets (Wang et. al. 2015).

The collocate clouds can be also used, which indicates the frequency of correlations of labels through character size and the correlations through colours. A key term is selected and the others following and proceeding it in different texts can be selected, giving a cloud (Savaşan & Diri, 2011).

In label clouds there are some weakness, including no fluency of visuals, ignorance of medium and small words, and limitation of evaluation to higher sizes (Hearst, 2008). However, label clouds are one of the most influential techniques in visualizing the data (Donath, 2002, Hearst & Rosner 2008).

Wordle, Tagul, Tagxedo, Jason Davies' Word Cloud Generator, WordSift, WordItOut, TagCrowd, Yippy, WordMosaic, AbcYa, VocabGrabber are among the word cloud software. Microsoft Excel can also be employed to carry out cloud activities.

2. CONCLUSION

The efficiency of the scientific measurement can be achieved through the correctness of the structured data, which should be free of mistakes. The reliability of the findings is directly based on the quality of the data of the study.

Text mining will be easier in the future, because Web 3.0 technologies, which are also called semantic web, will process the texts and objects stored in Internet. Therefore, many text mining procedures will be carried out before the placement of the texts into Internet, having ontologies, word and conceptual maps.

The major goal of the semantic web is to develop standards and technologies, presenting the well-defined and correlated data, which can be read and understood by machines (Arslan 2011).

The Web 3.0 technology has assumed the role of an assistant who knows many things. Now search engines can only give the web sites which cover the key word. However, the semantic web will interpret the websites and make it possible for us to access the correct websites in a short time.

The Web ontology language identifies the common and shared vocabulary or terminology used as standards. The Resource Description Framework (RDF) will put the vocabulary hierarchy and the correlations of objects into files.

REFERENCES

Akpınar, H., (2000), Veri Tabanlarında Bilgi Keşfi ve Veri Madenciliği. İ.Ü. İşletme Fakültesi Dergisi, 29(1), 1-22.

Arslan, A.A. (2011), *Türkçe Metinlerden Anlamsal Bilgi Çıkarımı İçin Bir Veri Madenciliği Uygulaması*, Başkent Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi (http://angora.baskent.edu.tr/acik_arsiv/dosya_oku.php?psn=2409&yn=352&dn=1)

Aydın, C.R., Erkan, A., Güngör, T. & Takçı, H., (2013). Sözlük Tabanlı Kavram Madenciliği: Türkçe için bir Uygulama, 30. Ulusal Bilişim Kurultayı, November 2013, Ankara. http://www.cmpe.boun.edu.tr/~gungort/papers/Sozluk%20Tabanli%20Kavram%20Madencili gi%20-%20Turkce%20icin%20bir%20Uygulama.pdf

Bilgin, T.T. ve Çamurcu, A.Y., (2008), Çok Boyutlu Veri Görselleştirme Teknikleri, Akademik Bilişim 2008, 30 Ocak - 01 Şubat. Çanakkale Onsekiz Mart Üniversitesi, Çanakkale, 107-112. http://ab.org.tr/ab08/kitap/Bildiriler/Bilgin_Camurcu_AB08.pdf

Blei, D., Ng, A. & Jordan, M. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3:993–1022. https://www.cs.princeton.edu/~blei/papers/BleiNgJordan2003.pdf

Cobo, M.J., López-Herrera, A.G., Herrera-Viedma, E. & Herrera, F. (2011). Science Mapping Software Tools: Review, Analysis, and Cooperative Study Among Tools, Journal of the American Society for Information Science and Technology, 62(7). 1382–1402, https://www.researchgate.net/publication/227733641_Science_Mapping_Software_Tools_Rev iew_Analysis_and_Cooperative_Study_Among_Tools

Cohen, K.B. & Hunter, L. (2008), Getting Started in Text Mining, PLoS Comput Biol. 4(1): e20. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2217579/

Çalış, K., Gazdağı, O. & Yıldız, O. (2013), Reklam İçerikli Epostaların Metin Madenciliği Yöntemleri ile Otomatik Tespiti, *Bilişim Teknolojileri Dergisi*, 6(1), 1-7. http://www.btd.gazi.edu.tr/article/viewFile/1041000157/pdf

De Bellis, N.(2009). *Bibliometrics and citation analysis: from the Science citation index to cybermetrics*. Scarecrow Press.

Dolgun, M.Ö.,Özdemir, T.G. & Oğuz, D. (2009), Veri madenciliğinde yapısal olmayan verinin analizi: Metin ve web madenciliği, *İstatistikçiler Dergisi*, 2, 48-58. http://dergipark.ulakbim.gov.tr/jssa/article/download/5000047710/5000045038

Donath, J., 2002, A semantic approach to visualizing online conversations, *Communications of the ACM*, 45(4). 45-49. DOI:10.1145/505248.505271

Friendly, M. (2006). A Brief History of Data Visualization, C. Chen, W. Hardle & A. Unwin (Eds), Handbook of Computational Statistics: Data Visualization. Springer. 1-43. http://www.datavis.ca/papers/hbook.pdf

He, Q. (1999). Knowledge Discovery Through Co-Word Analysis, *Library Trends*, 48(1). 133-159. https://www.ideals.illinois.edu/bitstream/handle/2142/8267/librarytrendsv48i1i_opt.pdf?sequ

Hearst, M. (2009), What is text mining, *http://www.sims.berkeley.edu/~hearst/textmining.html*

Hearst, M.A. & Rosner, D., (2008). Tag Clouds: Data Analysis Tool or Social Signaller?, *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual*, 7-10 Ocak 2008, 160-160. *DOI: 10.1109/HICSS.2008.422*

Hearst, M.A. (2008). What's Up With Tag Clouds, *Visual Business Intelligence Newsletter*, Mayıs. 1-5. https://www.perceptualedge.com/articles/guests/whats_up_with_tag_clouds.pdf

Hsieh, H.-F. & Shannon, S.E. (2005), Three Approaches to Qualitative Content Analysis, *Qualitative Health Research*, 15(9), 1277-1288. DOI: 10.1177/1049732305276687

Puri, S. (2011). A Fuzzy Similarity Based Concept Mining Model for Text Classification. Text Document Categorization Based on Fuzzy Similarity Analyzer and Support Vector Machine Classifier, International Journal of Advanced Computer Science and Applications (IJACSA), 2(11), 115-121. https://arxiv.org/ftp/arxiv/papers/1204/1204.2061.pdf

Savaş, S., Topaloğlu, N. & Yılmaz, M. (2012). Veri Madenciliği Ve Türkiye'deki Uygulama Örnekleri, *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, 11(21). 1-23. http://www.ticaret.edu.tr/uploads/yayin/fen21_dosyalar/1-23.pdf

Savaşan, S. & Diri, B. (2011). Automatic Tag Cloud Generation from Turkish Contents, Mühendislik ve Fen Bilimleri Dergisi Sigma 29, 156-169. http://eds.yildiz.edu.tr/ajaxtool/GetArticleByPublishedArticleId?PublishedArticleId=1867

Shi, C., Verhagen, M. & Pustejovsky, J. (2014). A Conceptual Framework of Online Natural Language Processing Pipeline Application, Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT, Dublin, Ireland. 53–59. http://www.aclweb.org/anthology/W14-5206

Şentürk, A. (2006). *Veri Madenciliği: Kavram ve Teknikler*, 2. Bursa: Ekin Yay.

Tan, A.H., Yu, P.S. (2004), Guest Editorial: Text and Web Mining, *Applied Intelligence 18, 239-241.*

Wang, L., Wang, G. & Alexander, C.A. (2015). Big Data and Visualization: Methods, Challenges and Technology Progress, Digital Technologies, 1(1), 33-38. http://pubs.sciepub.com/dt/1/1/7/