Tehseen, J Inform Tech Softw Eng 2018, 8:3
DOI: 10.4172/2165-7866.1000241

Journal of
Information Technology & Software Engineering

Review Article Open Access

Semantic Information Retrieval: A Survey

Rabia Tehseen*

University of Management and Technology, Lahor, Pakistan

Abstract

Semantic analysis is an ongoing research area. This survey paper deals with the latest updates in the field of text based semantic analysis. Various domains have been analyzed through this survey paper. The aim of this survey is to provide a complete picture and a brief description of the recent development in the text based aspect of the semantic information retrieval. The main contribution of this paper is to provide a highly sophisticated categorization of the large number of recent articles; in order to show the recent trends of research in the text based semantic analysis.

Keywords: Social media; Linguistics; Disaster; Encyclopedia

Introduction

Semantic analysis is a vector representation of text in words or documents that use document corpus as a knowledge base. In linguistics, semantic analysis is a process of relating syntactic structures, from the level of phrases, clauses, sentences and paragraphs to the level of writing as a whole, with their language independent meanings. In this paper, we have taken the text based aspect of the Semantic Analysis.

This survey paper provides a detailed overview of a text based aspect of a certain number of the recent articles that have been published in the field of semantic analysis. The author has classified the survey into four categories and has collected the articles accordingly. As the field has been growing every moment, a lot of developments occur along with time exploring new research areas and providing many recent domains to analyze the incorporated dataset. This survey paper provides an analysis of these domains to determine the trend of the researchers.

In this paper, a survey has been conducted. In which the author has collected thirty seven papers for the sake of semantic information retrieval. The articles were collected from the domains of social media, linguistics and knowledge, medicine and disaster management perspective. These domains have been further divided into sub categories as shown in Figure 1. In every branch, the authors have collected related articles that had been published within the year 2015 to 2018. Through this paper, the author has tried to find out that which are the active application areas or domains in semantic information retrieval.

The flow of the paper is as follows: in Section I, the author has discussed the categorization and sub-categorization of semantic information retrieval. In Section II, the author has discussed the proposed taxonomy and the reference table of the studied articles has been presented. The Section III has discussed all the papers that have been collected under each category and provides future trends. The Section IV discusses evaluation measures and data sets that had been used by the articles studied in the survey.

Contribution and Uniqueness

A large number of articles are presented in the field of semantic analysis every year. The amount of papers written in this domain is rapidly increasing. Hence, there has always been a need of such survey paper that may provide a summary of the recent articles to serve as an aid for the new researchers to get a detailed overview of the work that has already been done in this domain. This survey paper would serve as a starting step to facilitate researchers to set dimension for their work as they might know what has already been done and where there is still a need for more work.

Many papers have been presented in this year that has discussed semantic analysis techniques. Our work differs from the others in a way that we have focused on the application aspect of the semantic analysis. Moreover, the author has taken recent articles of the last three years including year 2015 till January 2018.

Section I: Categorization and sub-categorization

In this paper the author has tried to explore four different domains of semantic information retrieval.

- Social Media: It is the use of computer mediated technologies for the facilitation in the creation and sharing of information, ideas, carriers, interests and other forms of expressions via virtual communities and networks. The concepts of social media are at the top of the agenda of many business executives today. Consultants and decision makers try to find out ways in which they may make profitable use of open social applications. Social media have got an immense importance in regard of social networking and context sharing. There are many services that have emerged from business and social perspective. In this paper, the author has explored five different dimensions of social media, including news, blogs, OSN, reviews and forums. As shown in the taxonomy of Figure 1, the author has collected the information bout general news and financial news under the news node. Then, the articles conveying information through blogs have been listed. Information retrieved from Facebook, Twitter and Whatsapp has been represented under the heading of open social network (OSN). Customer reviews and their feedbacks have been collected under the node 'reviews'. For last information collected from the forums have been represented. Moreover, lifelogs have also been studied.

- Linguistics and knowledge domain includes articles concerning semantic similarity. Semantic similarity and plagiarism at word level, sentence level and document level had been discussed under linguistics. The author has collected the articles that have used content sharing services and encyclopedias in the heading of digital libraries. Moreover, the author has also represented the knowledge extracted from dictionaries in the same section.

***Corresponding author:** Rabia Tehseen, University of Management and Technology, Lahor, Pakistan, Tel: +92 42 35212801-10; E-mail: f2017288003@umt.edu.pk

Received August 25, 2018; **Accepted** September 10, 2018; **Published** September 15, 2018

Citation: Tehseen R (2018) Semantic Information Retrieval: A Survey. J Inform Tech Softw Eng 8: 241. doi: 10.4172/2165-7866.1000241

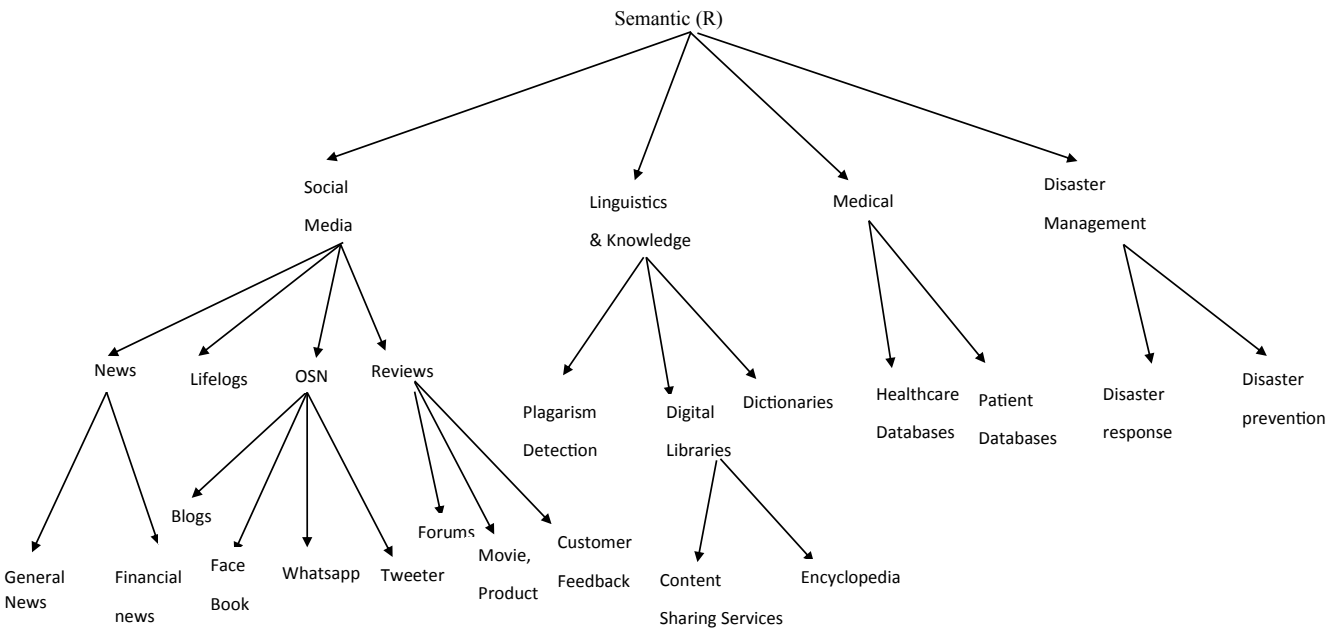

Figure 1: Proposed taxonomy.

- Medical and disaster management: medicine is an important field of information retrieval. In this survey the author has collected articles related to health care and patient records. Medical Information retrieval refers to methodologies and technologies that are used to access medical information through the process of information retrieval. This information is accessible from many sources, including general web, social media, journal articles and hospital records. Health related content is one of the most searched topics on the internet. This is highly emphasized domain of information retrieval research. The medical information is of the interest to a wide variety of users, including patients and their families, researchers and general practitioners with specific enterprise. The central issue in medical IR is the diversity of the users of these services. There are varying levels of information needs, medical knowledge and technical language skills. We have retrieved information from various health care databases and patient databases.

- Disaster response and prevention: One of the biggest challenge is to acquire information for getting prepared for natural disaster. Scientific data obtained from weather and seismic sensors are helpful in the prediction and preparedness of disaster. The purpose of disaster management is to save lives, preserve the environment and protect property and the economy. In the field of disaster management, we have covered disaster prevention/mitigation and response/recovery. All the papers related to disaster response and disaster prevention are collected under disaster management heading.

Table 1 shows the articles that have been studied from the respective domain.

Section II: Proposed taxonomy

The proposed taxonomy has been shown in Figure 1.

Discussion and Future Directions

Social media

As shown in Figure 1, the social media domain has further sub categories including news (general news and political news), blogs, OSN, Reviews, Forums and Lifelogs. Datasets used in these subcategories and evaluation measures taken have been discussed separately below.

News: The news industry has gone through a semantic shift in the past decade to include digital contents and a complete redefinition of how people consume news. We have collected articles from financial news and general news.

The author has compared news articles per section from two American newspapers. In this work, the news has been divided into three categories as hard news, soft news and general news. In order to segregate fake news, the news printed in the newspapers and online news from the social media (Facebook, Twitter) have been compared [1]. The author has collected data periodically over a period of two weeks from online sources and on the other hand 50,000 news articles published in the same time period in Guardian Newspaper and 11,000 articles from the New York Times newspaper have been compared. The information has been extracted about education, environment, lifestyles and entertainment. A similarity index has been developed on the basis of large volumes of commonly published articles.

The data to be analyzed include web documents, news articles, digital libraries and online forums [2]. In this work, the natural language text has been deconstructed according to the dependence between clauses. A semantic parser has been proposed. In this paper, three different text analyses have been conducted in which parsing techniques have been applied. To extract information about page count and snippets the proposed approach classified the web documents. The author has used REUTERS 21578 dataset. In future, the author aims to classify documents through random search based on the evolutionary algorithms. The author has studied the variance in the earnings of a firm by studying financial news [3]. The author has collected information about financial news through page count. Word pairs have been extracted through the application of multiple clustering patterns. The information extracted from financial news sources have been processed to calculate the rate of variance in the yearly earnings of a firm. The work has been evaluated by normalizing the earning of a firm. In to determine

Domain	Sub domains	Articles	Freq.	%
Social media	News	a4, a8, a13, a23, a26	22	60%
	Blogs	a3, a11, a27		
	OSN	a4, a6, a15, a18, a31, a28, a29, a30		
	Reviews	a6, a8, a15, a17, a19, a29, a30		
	Forums	a11, a13, a29		
	Lifelogs	a1, a24, a25		
Linguistics & Knowledge	Plagiarism detection	a2, a31, a32	18	49%
	Digital libraries	a7, a11, a12, a14, a15, a16, a20, a22		
	Dictionaries	a9, a5, a21		
Medical	Healthcare & Patient database	a10, a33, a34	3	8%
Disaster Management	Disaster response	a36, a37, a38	5	14%
	Disaster prevention	a34, a35		

Table 1: Reference Table of articles studied.

the sock market trend, stock exchange returns had been compared with the financial disclosure of the firms [3]. The author has calculated document level semantic relationship and have analyzed the data using their own proposed algorithm based on rhetorical structure theory (RST). The author has further explored various types of relationships of financial disclosure corpus and compared the results with the CODA parallel corpus (a benchmark corpus). The accuracy of the proposed algorithm has been predicted using machine learning engineering approaches. In future, the author aims to improve the performance of the proposed algorithm by exploring the sub level of RST tree.

To protect the financial information of a firm from competitors [4], the author has suggested a tradeoff in the information needs of a firm under security rules. They tied together elements from different literature standards and expressed the tradeoff in multiple needs of the firm. These needs, including balancing capital needs, liquid needs, pricing needs and security needs. The author urges that the firm's information can be secured from its rivals and its capital market share can be improved by making a tradeoff among firm's needs with the application of security rules.

Life logs: The author has focused on retrieving information related to life logging. The author has collected visual lifelogs using physical devices like cameras to capture localities and other social activities of a specific life logger [5]. Moreover, to represent real world information, the author has also recorded realistic topics experienced by the lifelogger. In this paper, three basic structures for information retrieval have been suggested including domain representation documents, application query topics and relevance judgement. By the application of these structures, the author has compared the experiences of social life and digital life of the life logger (Table 2).

The author has retrieved information from personal lifelog data. The NTCIR -13 lifelog core task has been reviewed [6]. Overall lifelogging personal experiences have been divided into sub tasks. For searching and retrieval of specific data Life logging Semantic Access subtask (LSAT) has been used. Lifelog even segmentation subtask has been used for knowledge mining and information retrieval from lifelogs.

Online Social Network: The researcher has worked on determining the effects of using social media applications like Facebook on female's mood, body posture and appearance [7]. An experiment has been conducted with a sample of 112 females and was asked to browse Facebook and a printed magazine for 10 minutes and then the changes in their moods have been noticed. The author has observed a clear fluctuation in the mood of the participant while using Facebook as compared to a fashion magazine. The author has performed the contextual analysis of various products from top brands through content sharing services (blogs) of social media [8]. The author urges

Number of Lifeloggers	3
Size of Collection (GB)	18.18GB
Size of Collection (Images)	88,124 Images
The size of a Collection (Long-Stay Semantic location)	130 Locations
The size of Collection (Visual Concept Metadata)	825 MB
The size of Collection (Visual Concept Detected)	1,000
Number of known item (LSAT) Topics	40
Number of Insight Analytics (LIT) Topics	10

Table 2: Statistics of NTCIR-12 Lifelog data [4].

that the identified brand must have at least a blog, external social network, Microsite, Microblog, a photo sharing app, mobile apps, social bookmarking, social games, virtual worlds, discussion forms or Wikis. The author has extracted information to identify the strategy or channel which is more successfully showing the consumer engagement with various brands using open social media applications. Various techniques have been used to attract customers like experience sharing, picture posting, exclusive messaging, voting and feedback.

The author has explained the effect of increased use of open social media applications including Facebook and Whatsapp on human behaviors [9]. A survey to analyze the effect of always remaining connected through online social media applications on human behavior has been conducted. The researcher has analyzed the navigational tracks of the users on Facebook, instant messaging and Twitter. A framework of ten human needs has been designed through the application of quantitative and qualitative methods and it has been inferred that Whatsapp is characterized for offering new opportunities to communicate whereas Facebook is a powerful life logging tool. A comparison of social media application and f 2 f communication has been presented by the Grefenstette and Muchemi [10]. The author has primarily retrieved information about the drastic change evolved due to technological growth in computing and the communication services offered by social media (Facebook, Whatsapp and blogs) in comparison with using traditional technologies like telephones, fax, e-mails. The growth rate in the number of users connected via open social media applications has been analyzed and it has been investigated that Facebook (launched in 2004) has more than 1.4 billion users and Twitter (launched in 2006) has 288 million. By analyzing the increase in the number of users it has been inferred that online social media applications are offering more flexibility due to their general purpose nature to the users and therefore, are more popular for communication than traditional ways of communication (Table 3).

Reviews: In order to perform concept level semantic analysis, data about different features from product reviews (including books, DVDs, and electronics) has been extracted [11]. To minimize redundancy and maximize relevance from the extracted data a concept extraction model

Service	Accounts
Facrbook	1,415
QQ	829
Whatsapp	700
Qzone	629
WeChat	468
Linkdin	347
Skype	300
Google+	300
Instagram	300
Baidu Tieba	300
Twitter	288
Viber	236
Tubmler	230
Snapchat	200
Line	181
Sina Weibo	167
Source: Statista.com as of March 2015	

Table 3: Loading social media services wordwide by active user accounts, Millions.

has been proposed through feature selection technique. The proposed model analyzes the extracted data using dependency based parsing techniques. The proposed model has been evaluated on a benchmark movie review dataset from Cornell University. In future, the author aims to explore more useful dependency relationships in product reviews by using concept net ontology.

The degree of content relatedness of the social web has been measured through explicit semantic analysis [12]. A similarity measure has been defined to estimate closeness among various content objects from social media. The researcher has classified the values of similar and dissimilar objects in a grid of 0 and negative numbers. The results have shown the rate of similar content available on open social media applications. The author has proposed "social extended semantic analysis" algorithm, which was based on explicit semantic analysis. The datasets of Facebook and Whatsapp dataset to analyze the proposed algorithm. The author aims to perform further experiments on large datasets, so that multimedia characteristics can also be examined through the proposed algorithm.

The author has analyzed the time varying response of social media text by capturing tweets from Twitter on monthly basis [13]. Topic based semantic analysis has been performed. The author has created a new n-gram corpus that has been delivered from 1.65 billion comments from online social media. An algorithm "translated latent semantic index" based on LSA have been created and used to generate their own corpus named as REDDIT. The work has been evaluated through an experiment on Reddit and Twitter datasets. In future, the author aims to use this corpus to study linguistic patterns of the social media text.

Customer feedback has been categorized into positive, negative and neutral feedbacks [14]. The researcher has provided a summarized review based on the customer feedback. Semantic analysis of reviews captured from the customer's feedback through mobile web services of the online social network has been performed. The work has been evaluated on Make My Trip and Twitter dataset to filter unwanted materials by tagging two separate parts of speech and writing summaries using rule based and supervised machine learning techniques. The author tends to improve the accuracy of their proposed technique in future.

The author has analyzed the human feedback regarding semantic similarity using open access datasets like MEN and WS-353 [15]. The researcher has determined that the measure of semantic similarity can be further improved by using additional information like Wikipedia text, navigational information and tagging along with these datasets. In this paper, the relationship between dimensionality and semantic expressiveness of embedded tags has been represented through graphs. To propose a universally accepted model, the author has extracted data about semantic relatedness from various domains. In the proposed model, low dimensional and high dimensional dense vector representation methods are used to analyze the data from open sources like Wikipedia text, Wikipedia navigational traces and from tagging. The data set used for the evaluation included MEN, WS353, WIKINAV and Bib100. In future, the author aims to explore other graph embedding algorithms for tagging data.

Linguistics and knowledge

Plagiarism detection: The researcher has worked on cross language plagiarism detection between French, English and Spanish language by making clusters and determining the strong Pearson correlation [16]. The plagiarism problem due to the translation of an article from one language to the other has been discussed in this paper. The author has proposed a unique technique for semantic evaluation of the main text by using several language pairs. In the proposed technique, the similarity at three different granularity levels have been gauged including document level similarity, sentence level similarity and chunk level similarity. Highly strong correlation among languages and text units has been considered. Plagiarism detection and evaluation have been conducted at chunk level and at sentence level. The author has proved that if one method is efficient for one language pair then it would be equally efficient on another language pair. Clustering match and mismatch have been calculated for the clusters that are apparently similar to each other. The author has shown that the methods behave differently in cluster match and mismatch; although they seem to be similar apparently regarding performance. The Logger man algorithm has been used by the author.

The author has presented an overview of five different source extraction methods for plagiarism detection and copyright violence [17]. The author has performed information flow analysis using cross year evaluation techniques. The author has performed source retrieval in three different phases, including innovation, consolidation and production. In the innovation phase, the author has introduced new corpora, technologies and performance measures. In the second phase the feedback and results from the first phase had been analyzed. In the last phase, the task has been observed more considerately and new ideas had been incorporated for the sake of innovation. The author has established an experimental platform named TERA for plagiarism detection at various levels (Figure 2).

The researcher has used four different standard methods, including recall, granularity, average precision and F-measures for the detection of plagiarism [18]. The author compares the performance of these

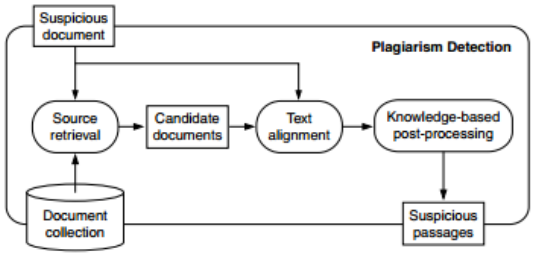

Figure 2: Generic retrieval process to detect plagiarism.

methods with other well-known methods. Moreover, the author has proposed a method to detect plagiarism by use word to word and document level interactions. It is an improved method in a sense that if a sentence has been used in different meanings than the queried text, then the proposed method has the ability to avoid the sentences having a high similarity index.

Digital libraries: Academic tests have been generated for the educational activities using automatic test generation system [19]. Data have been retrieved with respect to knowledge monitoring, term evaluation and topic wise. Moreover, it has been ensured that similarity index among the tests developed in the field of humanities and technical science have to be different. The author has used the classification templates to generate unique tests from the given 17 terms for a single person. The author has used AGTOaT algorithm to accomplish the task. The system proposed through this article has also been capable to generate unique tests in Russian language as well.

The researcher has collected sets of student summaries and worked to extract the data from quality summary sets by generating computerized summaries through proposed a crowd sourcing algorithm [20]. The proposed algorithm has been compared with seven other summary collection methods. The researcher has computed human summary scores and compared them with the scores of the expert good summaries and the crowd sourcing good summaries. The author has shown that in result, crowd sourcing summaries have obtained higher scores. The author has used Latent analysis corpora developed by the TASA organization for the evaluation of crowd sourcing semantic algorithms. The author has developed and evaluated crowd sourcing approach to automate scoring of summaries made by students. In future, the author aims to relate the text structure of the summaries to the performance of the proposed algorithm.

The authors have worked on a concept-concept association using the "see also" link graph of Wikipedia [21]. The main purpose of the article is to highlight the issue of sparsity among concepts and the restrictions of embedded text structures in concept-concept association. In order to handle these issues the author has proposed mined semantic analysis (MSA) method. Graph based approach and Mined semantic analysis (MSA) has been conducted on the data obtained from Wikipedia search index and the concept-concept association repository. The researchers have proved that the application of MSA has been a better approach as compared to the traditional method for clustering of semantic relatedness. The author has used the MC, RG, WS datasets for the evaluation of the proposed technique. The proposed MSA digs out the implicit concept-concept association, but in the future the author aims to extend his work from concept-concept explicit association.

Dictionaries: The author has analyzed the semantically rich information system [22]. Historical data with rich information about letters, words and synonyms of English language of the Oxford dictionary have been retrieved. The author has used a contextual disambiguation algorithm to develop a semantic trigger on the Oxford English dictionary for the annotation of lexical units. Language data have been triggered to annotate all the lexical units of the text through fine grained semantic categorization scheme provided by the Oxford English dictionary. In future, the author aims to make this tool more powerful to be able to handle corpus based studies.

The researchers [23] have classified the information about linguistics and have worked to design an algorithm to select terms from English dictionary that may serve as a base for differentiation between various topics from neural network systems. To classify information of linguistics, the author has developed a classification algorithm and has evaluated the proposed algorithm using R programming language. The proposed algorithm has also been capable to analyze synonyms for avoiding redundancy. The author has made web base classification mechanism and have presented an innovative idea to represent text document through string vectors instead of numeric ones. The newly developed system has the capability to classify a set of English text documents into a number of classes depending upon their contents. In future, the author intends to work for improvement in the accuracy of the system.

The authors have developed a semantic analyzer to analyze test results obtained from the automated test generation system for the students [24]. Lexicon based semantic algorithm has been used by the author to compare test results with the linguistic data obtained from dictionary, facts and common sense. An automatic test generating system has been proposed which is capable to generate indirect answers from the wide range of dictionary words. An interpreter has been developed to check the acceptance and rejection of the proposed answer from the dictionary.

Medical

Healthcare

The healthcare field is diverse and is certainly linked to the health delivery system that is run by both state and non-state authorities. People living in urban areas have an easy access to the best *healthcare* facilities whereas, in rural areas, accessibility to proper medical services is difficult.

Patient database

The authors have collected data about burnout patients from the patient database [25] and have determined the relationship between overall burnout cases with the newly developed scale for quality of care provided to them. The author has defined three distinct factors (Client-Centered Care, General Work Conscientiousness, and Low Errors) for self-reported cases, with good internal consistency. Burnout, particularly personally reported, and to some extent depersonalization have predicted the quality of Care provided to them.

The authors have collected cross sectional data from patient care database with the purpose to compare it with the British Thoracic asthma registry [26]. The author has retrieved the information about the impact of steroids and morbidity prevalence rate. Systematic steroid exposure and increase in the rate of morbidity prevalence has been correlated and their impact has been reported separately.

The author has studied dye-sensitized cells from patient documents [27]. The researcher has combined subject-action-object (SAO) approach of semantic analysis along with the morphological analysis. The author has presented a systematic approach for the identification of new technological opportunities. A five staged process has been proposed for analyzing patient documents to determine new opportunities. The author has used data about patient documents from Darwent innovations Index (A patient database). Author has used NLP techniques for the extraction of SAO structures to obtain information about the single patient. The author aims to automate the proposed system to establish morphology matrices in the future.

Disaster Management

Disaster prevention

The author has worked on urban emergency events. The author has performed spatial and semantic analysis by using the sensor based

model [28]. The main focus of the paper is on disaster detection, disaster response and control, rescue resource planning and scheduling and emergency commands. GIS and keyword based searching algorithm have been used to analyze data from OSN and blogs. The author has analyzed the proposed system on twitter, Welbo and Chinese micro-blog data set. The article analyzes the data in English as well as in Chinese language. The author aims to include public opinions and requirements with respect to spatial and temporal prospective to help government offer more effective assistance.

In the field of disaster management, social media has emerged as a valuable resource in handling the crisis situation like in the disasters caused by natural hazards. An analysis has been performed and an approach for the identification of relevant messages from social media has been proposed [29]. Relevant messages have been extracted by analyzing the relationship among three variables; including Geo referenced social media, geographic features and volunteer information.

Disaster response

Disaster response is an action taken in response to an unexpected and dangerous occurrence in an attempt to mitigate its impact on people or on the environment. Disaster situations can range from natural disaster to hazardous materials problems and transportation incidents. Disaster response may refer to the services provided by the government or rescue agencies to respond to emergencies.

The author has worked on the strategic decision making which is highly important for the successful disaster response [30]. The purpose of the paper is to develop an optimized model to minimize the risk to some extent that has already been exposed in certain areas. The aim is to establish a reliable framework with a choice to choose the location set from where facilities can be provided to the disaster area. The author has analyzed the information extracted through micro-blogging via text messaging on twitter for speedy disaster response in [31]. In this study, the author has compared the usefulness of label data obtained from a prior source with the unlabeled data from the current disaster, in order to calculate domain adaptation classifier for the target area. Experimental results have predicted that domain adaptation approach for labeled data stands superior to unlabeled data/ source data in disaster response.

The author has focused on the issue of ambulance routing for disaster response as medical aid is needed for the large amount of people [32]. The author has proposed two mathematical formulations to establish route plans with the minimum wait time. The author has analyzed different structural parameter to address the routing problem. The author has analyzed the data collected from the disaster area through aerial vehicles of the territory for the disaster response [33-36]. It has been observed that the spatial data collected by UASs are more sophisticated as compared to satellite data. The author has proposed hybrid crowd sourcing real time machine learning solution for efficient processing of aerial data with a huge volume (Table 4).

Dataset	Data Scope
DBNary [16]	News, Wikipedia
Twitter, Weibo, Chinese microblog Chen How [28]	Blogs
Reddit, Twitter [13]	Online Social Network (OSN)
Make My Trip, Twitter, OYO [14]	OSN, customer feedback
CODA, Financial news corpus [3]	Movie review, stock market, news
Oxford English dictionary [22]	Language data
Derwent innovations Index (patient database) [27]	Patient database
Movie review dataset, Product review dataset [11]	Forum, blogs, social network, content sharing service
LSA corpus created by TASA [20]	Student summaries
REUTERS 21578 [2]	Web documents, news articles, digital libraries, online forum
Facebook dataset, Whats app dataset [12]	Social media, facebook, whats app
MC, RG, WS.WSS.WSR [21]	Wikipedia search index, concept-concept association repository
MEN, WS 353, WIKINAV, BIB100 [15]	Wikipedia text, tagging, navigational tracks
Twitter, Facebook [33]	Text fragments of OSN
Google news corpus, MEDLINE articles from world-cat [34]	Medical reviews
ROUGE [36]	Student text summaries
English CELEX data, German Drive Base data [35]	Text summaries
Wikipedia Probase data [36]	Benchmark dataset and English drive based data
News article repository [1]	Data collected from the New York Times & Guardian Newspaper corpus
NTCIR [5]	Lifelog Test Collection
Multimodal lifelog data [6]	Data is collected over 90 days by two lifeloggers
Initial Public Offering files [4]	
Facebook, Twitter, You tube dataset [7-9]	OSN, blogs and content sharing service
Webis-TRC-2012 [17]	
PAN-PC-10,PAN-PC-11 [18]	
Optimal Patient Care Research Database (OPCRD) 16, British Thoracic Society (BTS) Difficult Asthma Registry17 [26]	Asthma Patient records
TWITTER DATA, HydroSHEDS [25]	Open social networking data, tagging
Twitter dataset [31]	Tagging
Patient database [32]	Patient records

Table 4: Datasets.

Conclusion

This survey paper has illustrated most of the recent updates in the field of semantic analysis. Text based aspect of the semantic analysis had been presented in this survey paper. Fifty one article from the year 2015 till 2018 had been analyzed, categorized and summarized. This survey has also illustrated the multi lingual work of various researchers. Moreover, the analysis of the studied domains through these papers has determined the trend of the researchers in recent years, in the field of semantic information retrieval.

References

1. Gurrin C, Joho H, Hopfgartner F, Zhou L, Gupta R, et al. (2017) Overview of NTCIR-13 lifelog-2 task. NTCIR 6-11.
2. Ntalianis K, Otterbacher J, Mastorakis N (2017) Content relatedness in the social web based on social explicit semantic analysis. In AIP Conference Proceedings 1836: 020068.
3. Markle-Huß J, Feuerriegel S, Prendinger H (2017) Improving sentiment analysis with document-level semantic relationships from rhetoric discourse structures. In Proceedings of the 50th Hawaii International Conference on System Sciences 1142-1150.
4. Fardouly J, Diedrichs PC, Vartanian LR, Halliwell E (2015) Social comparisons on social media: The impact of Facebook on young women's body image concerns and mood. Body Image 13: 38-45.
5. Boone AL, Floros IV, Johnson SA (2016) Redacting proprietary information at the initial public offering. Journal of Financial Economics 120: 102-123.
6. Ashley C, Tuten T (2015) Creative strategies in social media marketing: An exploratory study of branded social content and consumer engagement. Psychology & Marketing 32: 15-27.
7. Obar JA, Wildman SS (2015) Social media definition and the governance challenge: An introduction to the special issue. Telecommunications Policy 39: 745-750.
8. Karapanos E, Teixeira P, Gouveia R (2016) Need fulfillment and experiences on social media: A case on Facebook and WhatsApp. Computers in Human Behavior 55: 888-897.
9. Cappellato L, Ferro N, Jones G, San Juan E (2015) Source retrieval for plagiarism detection from large web corpora: recent approaches. Sematic Scholar 1391.
10. Grefenstette G, Muchemi L (2016) On the Place of Text Data in Lifelogs, and Text Analysis via Semantic Facets. arXiv preprint arXiv:1606.02440.
11. Chinniyan K, Gangadharan, S, Sabanaikam K (2017) Semantic Similarity based Web Document Classification Using Support Vector Machine. IAJIT 14: 285-292.
12. Ntalianis K, Otterbacher J, Mastorakis N (2017) Content relatedness in the social web based on social explicit semantic analysis. In AIP Conference Proceedings 1836: 020068.
13. Dang A, Moh'd A, Islam A, Minghim R, Smit M, et al. (2016) Reddit Temporal N-gram Corpus and its Applications on Paraphrase and Semantic Similarity in Social Media using a Topic-based Latent Semantic Analysis. InProceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers 3553-3564.
14. Fernandes R, Rio D'Souza GL (2017) Semantic Analysis of Reviews Provided by Mobile Web Services Using Rule Based and Supervised Machine Learning Techniques. International Journal of Applied Engineering Research 12: 12637-12644.
15. Li H, Cai Z, Graesser AC (2017) Computerized summary scoring: crowdsourcing-based latent semantic analysis. Behav Res Methods 1-18.
16. Ferrero J, Besacier L, Schwab D, Agnes F (2017) Deep Investigation of Cross-Language Plagiarism Detection Methods. arXiv preprint arXiv:1705.08828.
17. Abdi A, Idris N, Alguliyev RM, Aliguliyev RM (2015) PDLK: Plagiarism detection using linguistic knowledge. Expert Systems with Applications 42: 8936-8946.
18. Sweeney J, Patterson CC, Menzies-Gow A, Niven RM, Mansur AH, et al. (2016) Comorbidity in severe asthma requiring systemic corticosteroid therapy: cross-sectional data from the Optimum Patient Care Research Database and the British Thoracic Difficult Asthma Registry. Thorax 71: 334-346.
19. Arzhakov AV, Silnov DS (2016) New approach to designing an educational automated test generation system based on text analysis. ARPN Journal of Engineering and Applied Sciences 11: 2993-2997.
20. Kakade A, Dhumal K, Das S, Jain S, Ranjan NM (2017) A Neural Network Approach for Text Document Classification and Semantic Text Analytics. Journal of Data Mining and Management, Volume 2.
21. Campr M, Ježek K (2015) Comparing semantic models for evaluating automatic document summarization. In International Conference on Text, Speech, and Dialogue 9302: 252-260.
22. Wang X, Ma P, Huang Y, Guo J, Zhu D, et al. (2017) Combining SAO semantic analysis and morphology analysis to identify technology opportunities. Scientometrics 111: 3-24.
23. Shalaby WAF, Zadrozny W (2017) Semantic Representation Using Explicit Concept Space Models. InAAAI 4983-4984.
24. Boguslavsky I, Dikonov V, Frolova T, Iomdin L, Lazurski A, et al. (2016) Plausible Expectations-Based Inference for Semantic Analysis. I Int'l Conf. Artificial Intelligence 477-483.
25. Salyers MP, Fukui S, Rollins AL, Firmin R, Gearhart T, et al. (2015) Burnout and self-reported quality of care in community mental health. Adm Policy Ment Health 42: 61-69.
26. De Albuquerque JP, Herfort B, Brenning A, Zipf A (2015) A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. International Journal of Geographical Information Science 29: 667-689.
27. Agarwal B, Poria S, Mittal N, Gelbukh A, Hussain A (2015) Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach. Cognitive Computation 7: 487-499.
28. Xu Z, Zhang H, Sugumaran V, Choo KKR, Mei L, et al. (2016) Participatory sensing-based semantic and spatial analysis of urban emergency events using mobile social media. EURASIP Journal on Wireless Communications and Networking 2016:44.
29. Akgün İ, Gümüşbuğa F, Tansel B (2015) Risk based facility location by using fault tree analysis in disaster management. Omega 52: 168-179.
30. Li H, Guevara N, Herndon N, Caragea D, Neppalli K, et al. (2015) Twitter Mining for Disaster Response: A Domain Adaptation Approach. InISCRAM.
31. Talarico L, Meisel F, Sörensen K (2015) Ambulance routing for disaster response with patient groups. Computers & Operations Research 56: 120-133.
32. Ofli F, Meier P, Imran M, Castillo C, Tuia D, et al. (2016) Combining human computing and machine learning to make sense of big (aerial) data for disaster response. Big Data 4: 47-59.
33. Cotterell R, Schütze H (2017) Joint semantic synthesis and morphological analysis of the derived word. arXiv preprint arXiv:1701.00946.
34. Patel C, Gadhavi M (2017) A Model for Document classification using Kernel Discriminant Analysis (KDA) and semantic analysis. International Journal of Advanced Research in Computer Science 8: 783-786.
35. Bastos MT (2015) Shares, pins, and tweets: News readership from daily papers to social media. Journalism Studies 16: 305-325.
36. Gurrin C, Joho H, Hopfgartner F, Zhou L, Albatal R (2016) Ntcir lifelog: The first test collection for lifelog research. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval 705-708.