A Literature Survey on Web Content Mining

V. David Martin
Research Scholar, Department of Computer Science
Periyar E. V. R College (Autonomous)
Trichy, Tamilnadu, India
davidmartin17.cs@gmail.com
dmphd2016@gmail.com

Dr. T. N. Ravi
Assistant Professor, Department of Computer Science
Periyar E. V. R College (Autonomous)
Trichy, Tamilnadu, India
proftnravi@gmail.com

Abstract**—**Web is an accumulation of inter related documents on one or more web servers while web mining implies extricating important data from web databases. Web mining is one of the data mining spaces where data mining methods are utilized for extricating data from the web servers. The web information incorporates site pages, web links, questions on the web and web logs. Web mining is utilized to comprehend the client behavior, assess a specific site in view of the data which is stored in web log documents. Web mining is assessed by utilizing data mining strategies, specifically Association Rules, Classification and Clustering. It has some helpful regions or applications, for example, Electronic trade, E-learning, E-government, E-arrangements, E-majority rules system, Electronic business, security, crime examination and computerized library. Recovering the required web page from the web productively and adequately becomes a challenging task since web is comprised of unstructured information, which conveys the substantial measure of data and increment the unpredictability of managing data from various web service providers. The accumulation of data turns out to be elusive, extract, channel or assess the significant data for the clients. In this paper, we have considered the essential ideas of web mining, classification, procedures and issues. Notwithstanding this, this paper likewise broke down the web mining research challenges.

***Keywords** - Web Mining, Web Content Mining, Web Usage Mining, Web Structure Mining, Classification, Applications, Research Issues.*

*****

I. INTRODUCTION

Web mining is the use of data mining strategy which is an unstructured or semi organized information and it naturally finds and extracts conceivably helpful and previously obscure data or learning from the web [1]. The noteworthy web mining applications are web outline, web look, web crawlers, data recovery, organize administration, Ecommerce, business and Artificial Intelligence, web commercial centers and web groups. Online business breaks the boundary of time and space when contrasted with the physical office business. Enormous organizations around the globe understand that e-business is not simply purchasing and offering over Internet, rather it enhances the proficiency to contend with different monsters in the market. This application incorporates the fleeting issues for the clients.

Web mining has three classifications namely to be specific, the web content mining, the web structure mining and the web usage mining. Every classification is having its own tools and algorithms. Web content mining is only the disclosure of significant data from web reports and these web archives may contain content, picture, hyperlinks, metadata and organized records. It is utilized to look at the data via web search tool or web bugs i.e. Google, Yahoo. It is the way toward recovering the helpful data from the web substance or web reports. Web structure mining is likewise a procedure of finding organized data from the sites. The structure of a diagram comprises of pages and hyperlinks where the web pages are considered as hubs and the hyperlinks are edges and these are associating between related pages. Web usage mining is likewise called as the web log mining. It mirrors the client's behavior which can get the important examples from one or more web areas [2].

Web mining process comprises of four imperative strides, they are, resources finding, information selection and pre-processing, speculation and investigation [1]. Resources finding is the procedure which is utilized to separate the information either from online or disconnected content resources. In information selection and preprocessing step, particular data from recovered web sources are naturally chosen and pre-processed. Amid speculation, information mining and machine learning methods are utilized to find general examples from individual sites and over numerous locales. Approval and understanding of the mined examples are done in investigation step [3][4]. Web mining is grouped into three distinct classes, they are, web content mining, web structure mining and web usage mining.

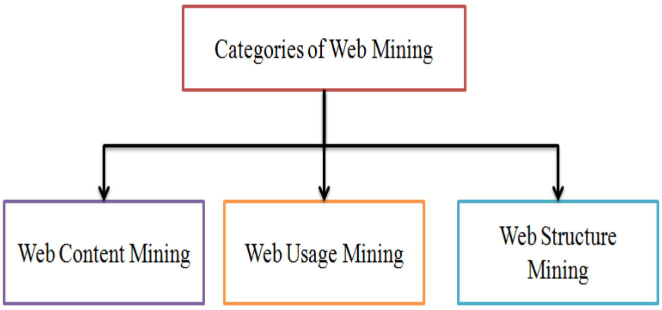

Figure 1 Categories of Content Mining

Web Mining is the utilization of data mining procedures to naturally find and concentrate data from web archives and services [5]. This territory of research is so gigantic today somewhat because of the interests of different research groups, the enormous development of data sources accessible on the web and the late enthusiasm for e-business. This marvel incompletely makes perplexity when we ask what constitutes Web Mining and when looking at research around there. The decaying web mining into these subtasks, to be specific:

- Resource Finding: The task of retrieving expected Web Documents

- Information Selection and Pre-Processing: Automatically selecting and preprocessing particular data from recovered Web Resources.
- Generalization: Automatically finds general examples at individual sites and in addition over different locales

II. RESEARCH ISSUES ON WEB MINING

The web is highly dynamic; lots of pages are included, redesigned and expelled each day and it handles tremendous arrangement of data subsequently there is an entry of numerous number of issues or problems. Typically, web information is high dimensional, restricted query interface, keyword arranged search and constrained customization to individual clients. Because of this, it is extremely hard to locate the important data from the web which may make new issues. Web mining systems are Association Rules, Clustering and Classification which are utilized to comprehend the client behavior, assess a specific site by utilizing conventional data mining parameters. Web mining procedure is separated into four stages; they are resource finding, information determination and pre-processing, speculation and investigation [6] [7]. Web analytics or web measurements are one of the critical difficulties in web mining. The estimation components are hits, online visits, visits or client sessions and locate the unique visitor routinely used to measure the client effect of different proposed changes. Organization archive and large institutions use information from the sites [8]. The principle issue is that, distinguishing and/or anticipating extortion exercises. The algorithms of web usage mining are more effective and exact. Yet, there is a test that must be contemplated. Web cleaning is the most critical process however information cleaning gets to be troublesome with regards to heterogeneous information [9]. Keeping up precision in ordering the information should be concentrated. Although numerous classification methods exist the nature of clustering is still a question to be answered.

- Web information sets can be substantial; it takes ten too many terabytes to store on the database.
- It can't mine on a solitary server so it needs substantial number of server.
- Proper organization of software and hardware to mine multi-terabyte information sets.
- Limited customization, constrained scope, and restricted inquiry interface to individual clients.
- Automated information cleaning.
- Over fitting and under fitting of information.
- Over Sampling of information
- Scaling up for high dimensional information
- Mining arrangement and time arrangement information
- Difficulty in finding important data
- Removing new information from the web

III. wEB CONTENT MINING

Web content mining information might be organized or unstructured/semi organized despite the fact that a lot of web is unstructured. It is the way toward recovering the data from the web into more organized structures and ordering the data to recover rapidly or discovering useful data from web reports or web servers. Web content mining incorporates the web archives which may comprise of content, html, mixed media reports i.e., pictures, sound, video and sound. The query result mining contains the web indexed lists. It might be a structure archives or unstructured reports.

Web content mining utilized numerous tools and algorithms, for example, Correlation Algorithm, Genetic Algorithm and Cluster Hierarchy Construction Algorithm (CHCA). Screen-scrapper, Ontology based tools; Mozenda, Web Info Extractor (WIE), web content extractor and mechanization anyplace are content mining tools. Cloud clients require to separate the data from the cloud gave by web servers can make utilization of the web mining. Case in point, Web communities can be maintained the data, for example, facebook. That is the clients of same field of intrigue can be gathered and they can impart through the system. Computerized library performs robotized reference ordering utilizing web mining strategies. E-services incorporate e-banking, web indexes, online closeouts, on-line information service, long range informal communication, e-learning, blog investigation, and recommendation system and personalization. This can be dissected for the clients and empower arrangement to the clients in light of their proposals [10]. It has two methodologies; they are (i) data view and (ii) Database View.

Table 1: Categories of Web Mining

	Web Mining			
	Web Content Mining		**Web Structure**	**Web Usage**
	DB View	***IR View***		
View of Data	Semi structured Web site as DB	Unstructured Semi structured	Links structure	Interactivity
Main Data	Hypertext Documents	Text documents Hypertext documents Hypertext Documents	Links structure	Server Logs Browser Logs
Representations	Edge-labeled graph (OEM) Relational	Bag of words, n-grams Terms, phrases Concepts or ontology Relational	Graph	Relational table Graph
Method	Proprietary algorithms ILP (Modified) association rules	TFIDF and variant Machine learning Statistical (including NLP)	Proprietary algorithms	Machine Learning Statistical (Modified) Association Rules
Application Categories	Finding frequent substructures Web site schema discovery	Categorization Clustering Finding Extraction Rules Finding Patterns in text User Modeling	Categorization Clustering	Site construction, adaptation, and management Marketing User Modeling

A. Research Issues on Web Content Mining

Web content mining has number of research issues since it can separate the data from the web internet searchers.

- Data/Information Extraction focus on extraction of structured information from website pages, for example, items and indexed lists.
- Schema matching and integration of web data. The web contains huge measure of information, every site acknowledge comparable data in an unexpected way. Comparative information revelation is an imperative issue with loads of practical applications.
- Opinion extraction from online sources i.e. client ensures items, forums, online journals and chat rooms. Mining opinions are of huge outcome for showcasing knowledge and item benchmarking.
- Automatically segmenting web pages and distinguishing clamor is an intriguing issue in web application. It couldn't have copyright notices, advertisements and navigation links. Subsequently, separating the fundamental substance of the site page is essential issue in web application [11].

IV. RELATED WORKS ON WEB CONTENT MINING

To play out any website assessment, web visitor's data assumes a critical part, so as to help this, numerous tools are accessible. Li, L,Zhang and C. Also, Zhang [12] communicated that Web Mining is a well-known procedure for dissecting website visitor's behavioral examples in e-service frameworks. Jian Pei,J. Han, B.Mortazavi and Hua Zhu [13] found that Web Log Mining helps in removing intriguing and valuable patterns from the Log File of the server. H. Tao Shen, Beng Chin OOi and Kian-Lee Tan [14] recommended that HTML records contain more number of pictures on the WWW. Such archives' containing important pictures guarantee a rich wellspring of pictures cluster for which query can be created. The archives which are very required by clients can be put close to the home page of the site. Manoj Manuja and Deepak Garg [15] proposed that the improvement of web mining methods, for example, web measurements and metrics, process mining and web service optimization and so forth will empower the force of WWW to be figured it out. Jing Wang and others [16] found that shortcoming of both utility and frequency can be overcome by General Utility Mining Model. Miller and Remington [17] uncovered that the structure of connected pages has definitive effect figure on the ease of use. Geeta and others [18] recommended that the quantity of pages at a specific level, the quantity of forward links and the quantity of in backward links to a specific page mirror the conduct of visitors to a particular page in the site. However Garofalakis [19] called attention to that the quantity of hit counts computed from Log File is an untrustworthy pointer of page fame. Geeta and others [20] recommended that the topology of the site assumes a vital part notwithstanding log record measurements to help clients to have snappy reaction. Jia-Ching and others [21] found that Web Usage Mining helps in finding web navigational examples chiefly to anticipate route and enhance site administration. Lee and others [22] demonstrated that the web behavioral examples can be used to enhance the plan of the site. These patterns also could help in enhancing the business intelligence.

Table 2: Literature Survey on Web Content Mining

Title	Representation of the Documents	Methods	Application
Survey on Unstructured Documents Types			
Ahonen, et al. [23]	Bag of words and word positions	Episode rules	Finding Keywords and Key pharses Discovering grammatical rules and collections
Billsus and Pazzani [24]	Bag of words	TFIDF Naïve Bayes	Text Classification
Cohen [25]	Relational	Propositional rule based system Inductive Logic Programming	Text classification
Dumais, et al. [26]	Bag of words Phrases	TFIDF Decision Trees Naïve Bayes Bayes Net Support Vector Machine	Text categorization
Feldman and Dagan [27]	Concept categories	Relative entropy	Finding patterns between concept distributions in textual data
Feldman, et al. [28]	Terms	Association rules	Finding patterns across terms in textual data
Frank, et al. [29]	Phrases and their positions	Naive Bayes	Extracting key phrases from text documents
Freitag and McCallum [30]	Bag of words	Hidden Markov Models	Learning extraction models
Hofmann [31]	Bag of words	Unsupervised statistical Method	Hierarchical clustering
Honkela, et al. [32]	Bag of words with ngrams	Self-Organizing Maps	Text and document clustering

Junker, et al. [33]	Relational	Inductive Logic Programming	Text categorization Learning extraction rules
Kargupta, et al. [34]	Bag of words with ngrams	Supervised Hierarchical Clustering Decision trees Statistical Analysis	Text classification and hierarchical Clustering
Nahm and Mooney [35]	Bag of words	Decision trees	Predicting (words) relationship
Nigam, et al. [36]	Bag of words	Maximum entropy	Text classification
Scott and Matwin [37]	Bag of words Phrases Hypernyms and synonyms	Rule based system	Text classification
Soderland [38]	Sentences, and clauses	Rule learning	Learning extraction rules
Weiss, et al. [39]	Bag of words	Boosted decision trees	Text categorization
Wiener, et al. [40]	Bag of words	Neural Networks Logistic Regression	Text categorization
Witten, et al. [41]	Named entity	Text Compression	Named entity classifier
Yang, et al. [42]	Bag of words and Phrases	Clustering Algorithms k-Nearest Neighbor Decision Tree	Event detection and tracking
Survey on Semi Structured Documents Types			
Craven, et al. [43]	Relational and ontology	Modified Naive Bayes Inductive Logic Programming	Hypertext classification Learning Web page relation Learning extraction rules
Crimmins, et al. [44]	Phrase, URLs, and meta information	Unsupervised and supervised classification algorithms	Hierarchical and graphical Classification Clustering
F¨urnkranz [45]	Bag of words and hyperlinks information	Rule learning	Hypertext classification
Joachims, et al. [46]	Bag of words and hyperlinks information	TFIDF Reinforcement learning	Hypertext prediction
Muslea, et al. [47]	Bag of words, tags, and word positions	Rule learning	Learning extraction rules
Shavlik and Eliassi-Rad [48]	Localized bag of words, and relational.	Neural networks with reinforcement learning	Hypertext (homepage) classification
Singh, et al. [49]	Concepts and Named entity	Modified association rule Classification algorithm	Finding patterns in semi-structured texts
Soderland [50]	Sentences, phrases, and named entity	Rule learning	Learning extraction rules

V. CONCLUSION

This paper has discussed about the research issues and challenges in web mining and also provided detailed review about the basic concepts of web mining, web content mining, structure mining, usage mining, tools, algorithms and types. Several open research issues and drawbacks which are exists in the current techniques are also discussed. This study and review would be helpful for researchers those who are doing their research in the domain of web mining.

REFERENCES

[1] Kumar, Shyam Nandan. "World towards Advance Web Mining: A Review." American Journal of Systems and Software 3.2 (2015): pp. 44-61.

[2] Feiran Huang, Jia Li, Jiaheng Lu, Tok Wang Ling, Zhaoan Dong, "PandaSearch: A Fine Grained Academic Search Engine for Research Documents", 2015 IEEE 31st International Conference on Data Engineering, 1408 – 1411.

[3] Jianshan Sun, Gang Wang, Xusen Cheng, Yelin Fu, "Mining Effective Text to Improve Social Media Item Recommendation", Elsevier-Information Processing and Management, Volume 51, Issue 4, July 2015, pp.444-457.

[4] P Ristoski, H Paulheim, "Semantic Web in Data Mining and Knowledge Discovery: A comprehensive Survey", Elsevier-Services and Agents on the World Wide Web, 2016, pp. 1-22.

[5] Abdullah Gok, Alec Waterworth, Philip Shapira, "Use of Web Mining in Studying Innovation", Scientometrics, January 2015, Volume 102, Issue 1, pp.653-671.

[6] Maria N. Moreno, Saddys Segrera, Vivian F. Lopez, Maria Dolores Munoz, Angel Luis Sanchez, "Web Mining based Framework for Solving Usual Problems in Recommender

Systes: A Case Study for Movies Recommendation", Elsevier-Neurocomputing 2015, 1-9.

[7] Khushbu Patel, Anurag Punde, Kavita Namdev, Rudra Gupta, Mohit Vyas, "Detailed Study of Web Mining Approaches - A Survey", International Journal of Engineering Sciences and Research Technology, February 2015, pp.23-30.

[8] Md. Zahid Hasan, Khawja Jakaria Ahmad Chisty and Nur-E-Zaman Ayshik, "Research Challenges in Web Data Mining", International Journal of Computer Science and Telecommunications Volume 3, Issue 7, July 2012.

[9] G. Dileep Kumar, Manohar Gosul, "Web Mining Research and Future Directions", Advances in Network Security and Applications, pp.489-496.

[10] Roma Yadav, S.R. Tandan, "Web Content Mining Applications and Methods-A Survey", Software Engineering and Technology, Volume 7, Number 6 2015, pp.164-175.

[11] Clemens Koltringer, Astrid Dickinger, "Analyzing Destination Branding and Image from Online Sources: A Web Content Mining Approach", Elsevier-Journal of Business Research, pp.1-8, 2015.

[12] Yang, Q. and Zhang, H. , Web-Log Mining for predictive Caching, IEEE Trans. Knowledge and Data Eng., 15(4), 2003,1050-1053.

[13] Jain Pei, Jiawei Han, Behzad Mortazavi_asl and Hua Zhu, Mining Access Patterns Efficiently from Web Logs, Pacific-Asia Conference on Knowledge Discovery and Data Mining(PAKDD'00), Kyoto, Japan, 2000, 396 407.

[14] Heng Tao Shen, Beng Chin Ooi and Kian_Lee Tan, Giving meanings to WWW, ACM SIGM Multimedia, L.A,2000, 39-47.

[15] Manoj Manuja and Deepak Garg, Semantic web mining of Un-structured Data: Challenges and Opportunities, International Journal of Engineering, 5(3) ,2011, 268-276.

[16] Jing Wang, Ying Liu, Lin Zhou, Yong Shi, and Xingquan Zhu, Pushing frequency constraint to utility Mining Model, ICCS Springer-Verlag Berlin Heidlberg, LNCS 4489, 2007, 685-692.

[17] Miller, C.S. and Remington, R. W. Implications for information Architecture , Human Computer Interaction, Journal IEEE Web Intelligence, 2004, 19(3), 225-271.

[18] Geeta.R.B, Shashikumar G. Totad & Prasad Reddy PVGD, Optimizing User's Access To Web Pages, International refereed Journal JooiJA, Transactions on World Wide Web-Spring, 2008, 8(1), 61-66.

[19] Garofalakis, Web Site Optimization Using Page Popularity, IEEE Internet Computing, 1999, 3(4), 22-29.

[20] Geeta.R.B, Shashikumar G. Totad & Prasad Reddy PVGD, Topological Frequency Utility Mining Model Springer International Conference, SocPros 12, 2011, 505-508.

[21] Jia-ching Ying, Vincent S. Tseng, Philip S. Yu IEEE International Conference on Data Mining workshops IEEE Computer Society, 2009.

[22] Y.S.Lee, S.J Yen, and M.C.Hsiegh. A Lattice-Based Framework for Interactively and Incrementally Mining web traversal patterns, International Journal of Web Infornation Systems, 2005. 197-207.

[23] H. Ahonen, O. Heinonen, M. Klemettinen, and A. Verkamo. Applying data mining techniques for descriptive phrase extraction in digital document collections. In Advances in Digital Libraries (ADL'98), 1998.

[24] D. Billsus and M. Pazzani. A hybrid user model for news story classification. In Proceedings of the Seventh International Conference on User Modeling (UM '99), 1999.

[25] W. W. Cohen. Learning to classify english text with ilp methods. In Advances in Inductive Logic Programming (Ed. L. De Raedt). IOS Press, 1995.

[26] S. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In Proceedings of the 1998 ACM 7th international conference on Information and knowledge management, pages 148–155, Washington United States, 1998.

[27] R. Feldman and I. Dagan. Knowledge discovery in textual databases (kdt). In Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95), pages 112–117, Montreal, Canada, 1995.

[28] R. Feldman, M. Fresko, Y. Kinar, Y. Lindell, O. Liphstat, M. Rajman, Y. Schler, and O. Zamir. Text mining at the term level. In Principles of Data Mining and Knowledge Discovery, Second European Symposium, PKDD '98, volume 1510 of Lecture Notes in Computer Science, pages 56–64. Springer, 1998.

[29] E. Frank, G. W. Paynter, I. H. Witten, C. Gutwin, and C. G. Nevill-Manning. Domain-specific keyphrase extraction. In Proceedings of 16th International Joint Conference on Artificial Intelligence IJCAI-99, pages 668–673, 1999.

[30] D. Freitag and A. McCallum. Information extraction with hmms and shrinkage. In Proceedings of the AAAI- 99 Workshop on Machine Learning for Information Extraction, 1999.

[31] T. Hofmann. The cluster-abstraction model: Unsupervised learning of topic hierarchies from text data. In Proceedings of 16th International Joint Conference on Artificial Intelligence IJCAI-99, pages 682–687, 1999.

[32] T. Honkela, S. Kaski, K. Lagus, and T. Kohonen. Websom - self-organizing maps of document collections. In Proc. of Workshop on Self-Organizing Maps (WSOM'97), pages 310–315, 1997.

[33] M. Junker, M. Sintek, and M. Rinck. Learning for text categorization and information extraction with ilp. In Proceedings of the Workshop on Learning Language in Logic, 1999.

[34] H. Kargupta, I. Hamzaoglu, and B. Stafford. Distributed data mining using an agent based architecture. In Proceedings of Knowledge Discovery And Data Mining, pages 211–214. AAAI Press, 1997.

[35] U. Y. Nahm and R. J. Mooney. A mutually beneficial integration of data mining and information extraction. In Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-00), 2000.

[36] K. Nigam, J. Lafferty, and A. McCallum. Using maximum entropy for text classification. In Proceedings of the International Joint Conference on Artificial Intelligence IJCAI-99 Workshop on Machine Learning for Information Filtering, pages 61–67, 1999.

[37] S. Scott and S. Matwin. Feature engineering for text classification. In Proceedings of the 16th International Conference on Machine Learning ICML-99, 1999.

[38] S. Soderland. Learning information extraction rules for semi-structured and free text. Machine Learning, 34(1-3):233–272, 1999.

[39] S. M. Weiss, C. Apt´e, F. Damerau, D. E. Johnson, F. J. Oles, T. Goetz, and T. Hampp. Maximizing text-mining performance. IEEE Intelligent Systems, 14(4):63–69, 1999.

[40] W. Wiener, J. Pedersen, and A. Weigend. A neural network approach to topic spotting. In Proceedings of the 4th Symposium on Document Analysis and Information Retrieval (SDAIR 95), pages 317–332, 1995.

[41] I. H. Witten, Z. Bray, M. Mahoui, and W. J. Teahan. Text mining: A new frontier for lossless compression. In Data Compression Conference, pages 198–207, 1999.

[42] Y. Yang, J. Carbonell, R. Brown, T. Pierce, B. T. Archibald, and X. Liu. Learning approaches for detecting and tracking news events. IEEE Intelligent Systems, 14(4):32–43, 1999.

[43] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery. Learning to extract symbolic knowledge from the world wide web. In Proceedings of the Fifteenth National Conference on Artificial Intellligence (AAAI98), pages 509–516, 1998.

[44] F. Crimmins, A. Smeaton, T. Dkaki, and J. Mothe. T´etrafusion: Information discovery on the internet. IEEE Intelligent Systems, 14(4):55–62, 1999.

[45] J. F¨urnkranz. Exploiting structural information for text classification on the www. In Advances in Intelligent Data Analysis, Third International Symposium, IDA-99, pages 487–498, 1999.

[46] T. Joachims, D. Freitag, and T. Mitchell. Webwatcher: A tour guide for the world wide web. In Proceedings of the International Joint Conference on Artificial Intelligence IJCAI-97, pages 770–777, 1997.

[47] I. Muslea, S. Minton, and C. Knoblock. Wrapper induction for semistructured, web-based information sources. In Proceedings of the Conference on Automatic Learning and Discovery CONALD-98, 1998.

[48] J. W. Shavlik and T. Eliassi-Rad. Intelligent agents for web-based tasks: An advice-taking approach. In Working Notes of the AAAI/ICML-98 Workshop on Learning for Text Categorization, pages 588–589, 1999.

[49] L. Singh, B. Chen, R. Haight, P. Scheuermann, and K. Aoki. A robust system architecture for mining semi-structured data. In Proceeding of The Fourth Int. Conference on Knowledge Discovery and Data Mining (KDD-98), pages 329–333, 1998.

[50] S. Soderland. Learning information extraction rules for semi-structured and free text. Machine Learning, 34(1-3):233–272, 1999.