A Critique Survey on Diverse Approaches of Web Content Mining

Varish P V
Assistant Professor
School of CSA
REVA University: varishpv@reva.edu.in
Lokesh C K
Assistant Professor
School of CSA
REVA University: lokeshck@reva.edu.in
Dr. Kavitha
Associate Professor
School of CSA
REVA University: kavitha@reva.edu.in

----------**ABSTRACT**----------
Web mining is used to find the different patterns in data by various categories like web usage mining, web structure mining and web content mining. The method used to gather data by web spiders and web search engines are known as web content mining. The formation of a website can be tartan by using web structure mining and we can test the data of a user's browser by using web usage mining. The web content mining is a second phase of web mining, which deals with extraction of images, graphs, text etc. The spotlight of this work is to present a brief survey on different techniques used in web content mining. We presented a brief review of different web content mining approaches like multimedia mining, unstructured mining, structured mining and semi-structured mining.

Keywords: Web Mining, Web Content Mining, Multimedia Mining, Web Crawlers, Summarization, Information Extraction.

I. INTRODUCTION

Web is a communication channel it provides information over the internet. Websites are created by using web pages and stored in the web. The requirement of information stored in web is keep growing day by day. Using web people are doing various tasks like shopping, online transactions, applying for jobs etc. The different patterns of information gathered from the web and web is processed by using web mining. Web mining also helps to advance the command of web search engine by identifying the web pages and classifying the web documents. Web mining is incredibly helpful to e-commerce and e-services websites.

Web Mining

Web mining is associate application, information mining techniques to search out information patterns from the online data. Web mining helps to improve the supremacy of web search engine by identifying the web pages and classifying the web documents. Web mining helps to evaluate performance of a website and also to realize customer activities.

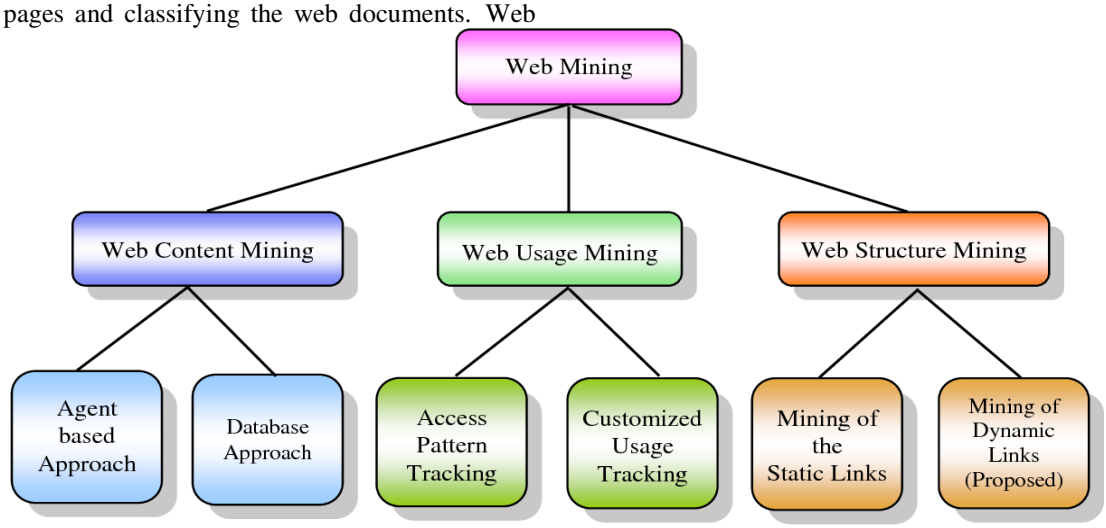

Fig 1: Tree structure of web mining

CATEGORIES OF WEB MINING

1. Web Content Mining

Web content mining consists of several types of data like, text, audio, video, image, hyperlinks etc.web content mining is the process of identifying and Fetching useful information, documents or data from the web. Web content mining can be used for Mining of versatile information from web pages. We can consider an example, if a web user need a suggestion to buy a product from web, the search engine provides the list of suggestions related to respective product.

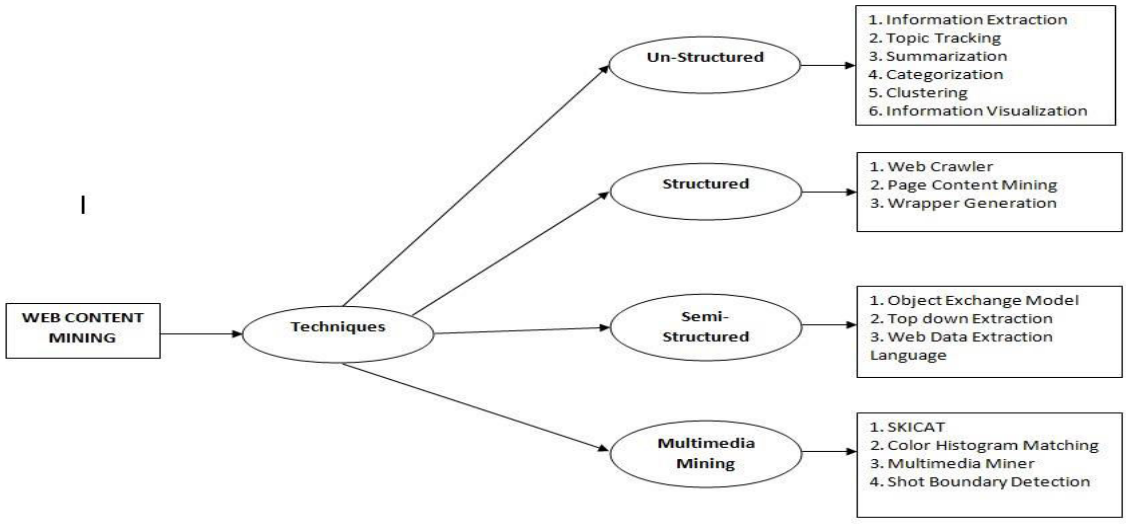

Fig 2: Techniques used for web content mining

2. Web Usage Mining

Web usage mining makes the sense of data generated by browsing sessions and behavior. Web usage mining keeps a record of web access called logs for the web browsing and helps to discover the user access patterns of web pages. It is the technique of extracting patterns and log records from the web data. With the growing popularity of the WWW, big size of data like user's address, requested URLs are automatically gathered by Web servers and stored in access log files. The techniques to determine and examine the web usage pattern are: Online Analytical Processing (OLAP) and Session& Analysis

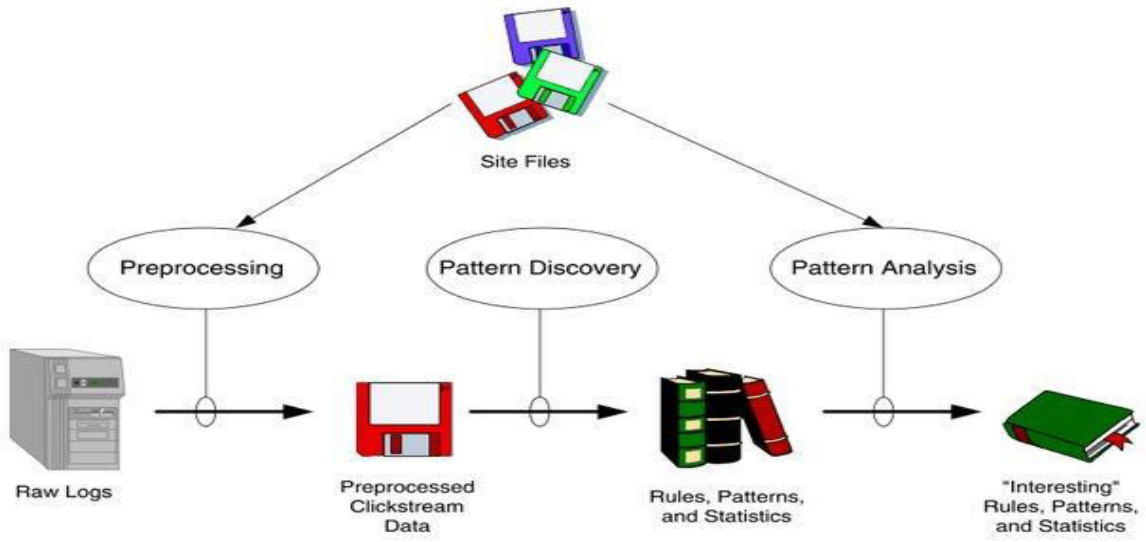

Fig 3: Log Structure of Web Usage Mining

3. Web Structure Mining

Using web structure mining we can study how the data in the websites are interlinked each other. Web Structure mining is helpful to recognize the web data either linked by information or direct link connections. The principle of structure mining is to construct the structural outline of website and comparable web pages.

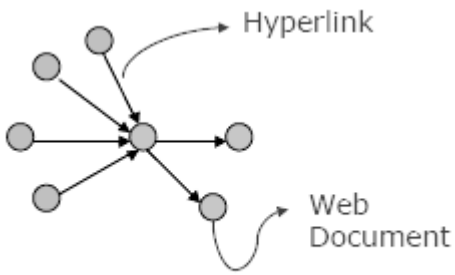

Fig 4: Web Graph Structure

II. WEB CONTENT MINING APPROACHES

The various techniques used for mining the content are:- unstructured, multimedia, semi-structured and structured mining.

Unstructured mining - The most content in web is in unstructured format. To extract text from unstructured format we use different techniques listed as follows: Categorization, Topic Tracking, Clustering, Information Extraction, Summarization, and Information Visualization.

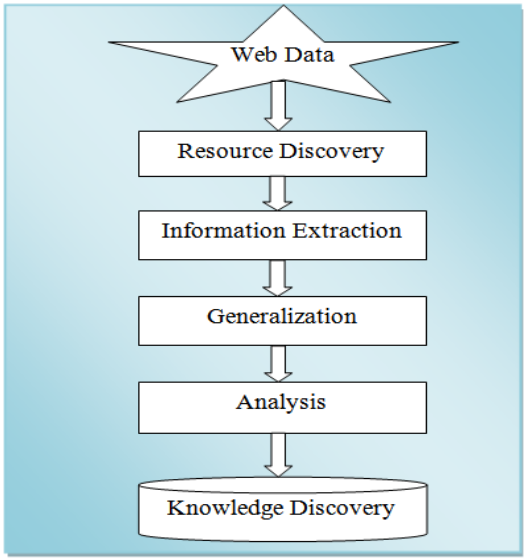

Fig 5: Unstructured Mining Techniques

Structured mining - The structured data is in the form of list, table, field set. The web crawlers are used to take data from client, search it and get good pages from the web. The external crawler takes user specified websites and expand the graph using newly found websites [4]

To extract text from structured format we can use the following techniques: Web Page Content Mining, Web Crawlers and Wrapper Generation [3].

Semi-structured mining-The partial structured and grammatical text from the web will come under this category, the various techniques used to extract this form of text are. To extract text from semi-structured format we can use the following techniques: Web Data Extraction Language, OEM - Object Exchange Model and Top down Extraction. [5]

Multi-media mining-The large amount of data in the form of text, audio, video is mined under this category. The various techniques used for multimedia mining are: Shot Boundary Detection, SKICAT, Multimedia Miner and Color Histogram Matching.[3][12]

III. LITERATURE REVIEW

Gandhi et.al [6] discussed in their research work about Information Extraction for Unstructured mining:–The Recourse Description Framework is used to maintain the accuracy of the system. The time factor of the searching information can be reduced by storing the description of research paper or articles. The meaning of paper or article keeps unchanged. Internet gives large amount of helpful information; most of the data is in the form of unstructured format. We can use two techniques to extract information from the web Extensible Markup Language and Recourse Description Framework (RDF); RDF extends the connecting structure of the web, distinct applications share semi-structured & structured data by using RDF. The author's proposed "Semantic Partitioning Object (SPO)" technique for rescue of research article or research paper data to achieve better accuracy and consistency.

The architecture for a simple information extraction system with various stages is shown in the below **Fig 6**. In first stage sentence segmenter will divide the raw text of the document into sentences and using tokenizer the sentence is further divided into words. Later every sentence is linked with tags called part-of-speech tags which will be useful for the next stage called entity detection; the interesting entity in each sentence is identified by using entity detection method. Lastly, to match likely relations between different entities in the text we use relation detection.

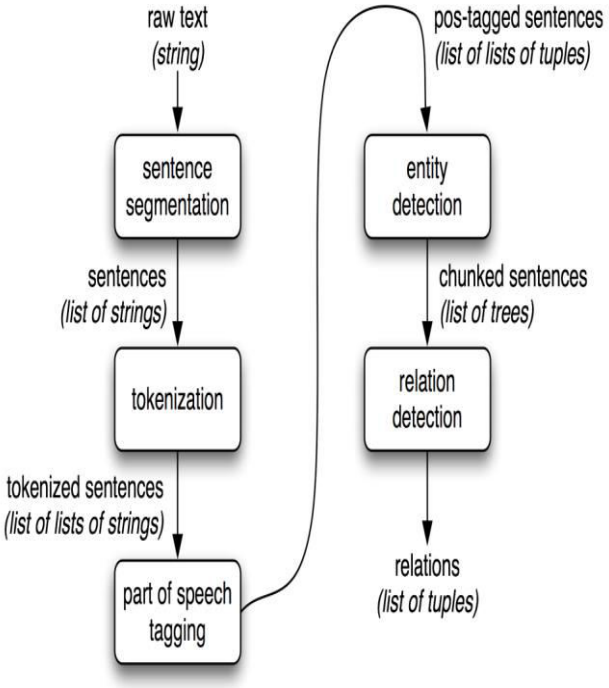

Fig 6: Information Extraction Architecture

Lee et.al [8]discussed in their research about topic tracking technique for news keyword extraction from the news websites. To track the topics from the news web sites we can use the important words called key-words. The meaning of each key-word cannot be stored manually as the online news data keeps growing rapidly. we require an mechanical procedure that extracts reserved words from news articles. Authors recommended Term Frequency–Inverse Document Frequency model (TF-IDF) module.

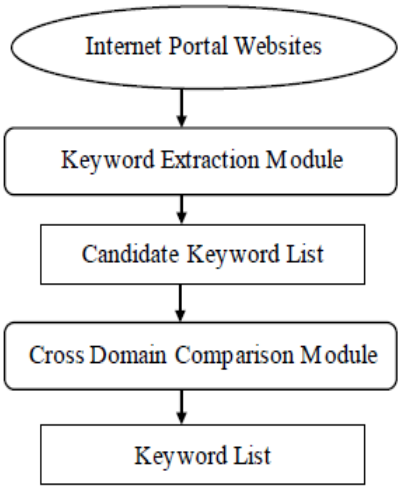

Fig 7: Keyword Exctraction

With the help web portals the HTML news pages are collected and by applying keyword pulling out method the candidate words will be taken out, finally keywords are extracted by cross-domain comparison module. The following database tables 'document', 'dictionary', 'term occur fact' and 'TF-IDF weight' are created relational database. In document table we store "downloaded news documents" Using document table the "nouns" are extracted. The document fact words are added to table konown as 'Term occur fact(TOF)'. By using TOF the TF-IDF weights for each word is calculated and the result are inserted to 'TF-IDF weight' table. Finally 'Candidate keyword list' for each news domain with words is ranked high by using TF-IDF weight table.

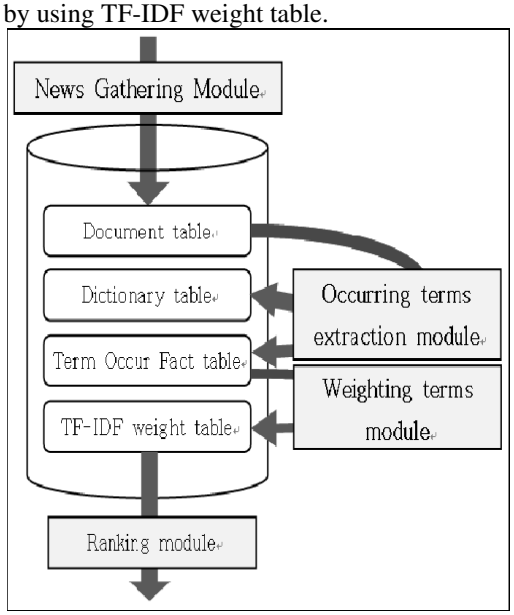

Fig 8: News Word Ranking Module[19]

Elfayoumy et.al[9] discussed in their research about unstructured text summarization techniques.The various algorithms are used to summerize the text from the web pages, each algorithm will have its own pros and cons. some algorithms works better for summarizing long documents where some are works better for short documents. In this research the authors presented text summarization algorithms with its strengths & weaknesses. Authors suggested that the collective approach will give better results for mixed types of documents.

The skill of abstracting key content from various sources of information is known as **S**ummarization. The public can take effective decisions in less time by using summarized text of updates from a news web sites, stock market, movie reviews from online sites etc [21]

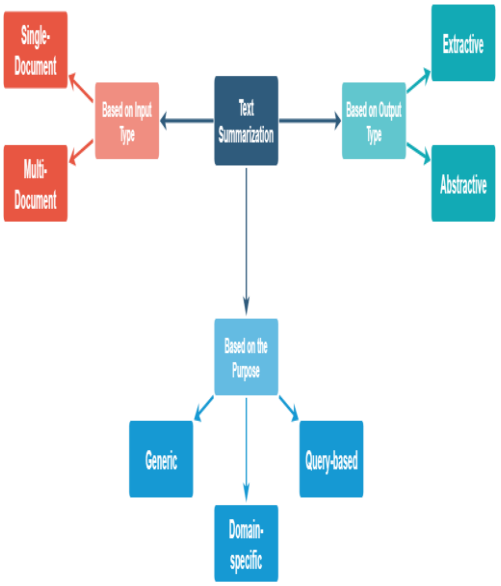

Fig 9: Text Summarization Method[19]

Subhendu Kumar Pani et.al[10] discussed in their research about web crawlers. Web spiders also named as web bots or robots or crawlers are the programs which is used to download the documents on the internet, web crawlers are most used in search engines. As the web-data is keep growing day by day it is difficult to download the content. The web crawelrs generate a replica of visited pages and stores all visited pages for final processing by a search engines. For information retrieval from web the web crawling is used as one of main module. crawling systematically traverse the World Wide Web(WWW). In this paper the authors proposed a crawling design that web crawlers collects web pages based on the users interest. Web crawling is apllied to improve the performance of the surf engines in World Wide Web.

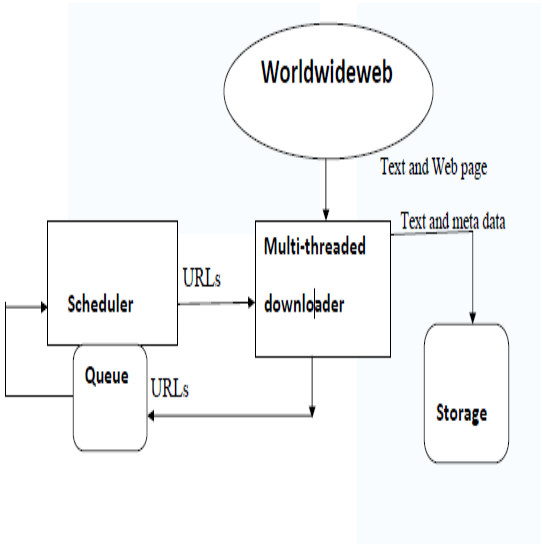

Fig 10: Web Crawler Architecture

Xia et.al[11] The authors in this journal proposes a automatic wrapper generation methods for content based websites. To discover identical structure of the input web pages like blogs, news sites or forums the tree alignment & transfer learning method is proposed. To get the consignment of different tag-matchings linear regression method is applied. Transfer learning method is adopted to find the fairly accurate content block. The wrapper generation method is applied to generate a wrapper once the web source change is detected.

Kotsiantis et.al[13] discussed in their research about multimedia mining.Advances in multimedia acquisition and storage technology have led to tremendous growth in verylarge and detailed multimedia databases. If these multimedia files are analyzed, useful information to users canbe revealed. Multimedia mining deals with the extraction of implicit knowledge, multimedia data relationships,or other patterns not explicitly stored in multimedia files. Multimedia mining is more than just an extension ofdata mining, as it is an interdisciplinary venture that draws upon expertise in computer vision,multimediaprocessing, multimedia retrieval, data mining, machine learning, database and artificial intelligence.

IV. CONCLUSION

The largse amount of knowledge and information is available from web. The web data is endlessly growing in volume and complexity with time so it is becoming difficult to mine the precious relevant information from internet. Thenumerous web content mining approaches or tools applied to take out relevant useful information and knowledge from the web page contents. This paper reviews the variousweb content mining approaches like Information Extraction, Topic Tracking, Summarization, Web crawlers, Wrapper Generation, Color Histogram and Multimedia Mining.

REFERENCES

1. Mughal, Muhammd Jawad Hamid. "Data Mining: Web Data Mining Techniques, Tools and Algorithms: An Overview." *International Journal of Advanced Computer Science and Applications* 9, no. 6 (2018).

2. Irfan, Shadab, and Subhajit Ghosh. "Web Mining for Information Retrieval." International Journal of Engineering Science 17277 (2018).

3. Mebrahtu, Andemariam, and Balu Srinivasulu. "Web Content Mining Techniques and Tools."

International Journal of Computer Science and Mobile Computing 6, no. 4 (2017).

5. Santosh Kumar Rath, Smaranika Mohapatra, And Jharana Paikaray. "Web Mining: A Tool for information retrieval from Online Marketing." *Internation Journal Of Advance Research And Innovative Ideas In Education* 2, no. 2 (2016) : 1329-1333.

6. Gandhi, Kalgi, and Nidhi Madia. "Information extraction from unstructured data using RDF." In 2016 International Conference on ICT in Business Industry & Government (ICTBIG), pp. 1-6. IEEE, 2016.

7. Gupta, Vishal, and Gurpreet S. Lehal. "A survey of text mining techniques and applications." Journal of emerging technologies in web intelligence 1, no. 1 (2009): 60-76.

8. Lee, Sungjick, and Han-joon Kim. "News keyword extraction for topic tracking." In 2008 Fourth International Conference on Networked Computing and Advanced Information Management, vol. 2, pp. 554-559. IEEE, 2008.

9. Elfayoumy, Sherif, and Jenny Thoppil. "A survey of unstructured text summarization techniques." *The International Journal of Advanced Computer Science and Applications* 5, no. 7 (2014): 149-54.

10. Subhendu Kumar Pani, Deepak Mohapatra and Bikram Keshari Ratha. UTKALUNIVERSITY, RCMA RCEM. "Integration of web mining and web crawler: Relevance and state of art." Integration 2, no. 03 (2010): 772-776.

11. Xia, Yingju, Yuhang Yang, Shu Zhang, and Hao Yu. "Automatic wrapper generation and maintenance." In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*. 2011.

12. Grundland, Mark, and Neil A. Dodgson. "Color histogram specification by histogram warping." In Color Imaging X: Processing, Hardcopy, and Applications, vol. 5667, pp. 610-622. International Society for Optics and Photonics, 2005.

13. Kotsiantis, S., D. Kanellopoulos, and P. Pintelas. "Multimedia mining." *WSEAS Transactions on Systems* 3, no. 10 (2004): 3263-3268.

14. Mohamad, Fatma Susilawati, Azizah Abdul Manaf, and Suriayati Chuprat. "Histogram matching for color detection: A preliminary study." In *2010 International Symposium on Information Technology*, vol. 3, pp. 1679-1684. IEEE, 2010.

15. Bharanipriya, V., and V. Kamakshi Prasad. "Web content mining tools: a comparative study." *International Journal of Information Technology and Knowledge Management* 4, no. 1 (2011): 211-215.

16. Johnson, Faustina, and Santosh Kumar Gupta. "Web content mining techniques: a survey." *International Journal of Computer Applications* 47, no. 11 (2012).

17. Malarvizhi, R., and K. Saraswathi. "Web Content Mining Techniques Tools & Algorithms–A Comprehensive Study." *International Journal of Computer Trends and Technology (IJCTT)* 4, no. 8 (2013): 2940-2945.

18. Kumar, Anurag, and Ravi Kumar Singh. "A Study on Web Structure Mining." *International Research Journal of Engineering and Technology (IRJET)* 4, no. 1 (2017): 715-720.

19. Sathya, S., and N. Rajendran. "A review on text mining techniques." *Int. J. Comput. Sci. Trends Technol* 3, no. 5 (2015): 274-284.

20. Azir, Mohd Amir Bin Mohd, and Kamsuriah Binti Ahmad. "Wrapper approaches for web data extraction: A review." In *2017 6th International Conference on Electrical Engineering and Informatics (ICEEI)*, pp. 1-6. IEEE, 2017.

21. Hahn, Udo, and Inderjeet Mani. "The challenges of automatic summarization" *Computer* 33, no. 11 (2000): 29-36.