Review of Structure and Unstructure Based Web Document Classification

Amit Rathore
CSE, Department, SIRTE, Bhopal, India
amit13aprl92@gmail.com

Dr. Kamlesh Namdev
CSE, Department, SIRTE, Bhopal, India
kamlesh.namdev@gmail.com

***Abstract*— the exponential growth of the web has raised the importance of web document classification in web classification, website pages from at least one sites are allocated to pre-characterized classes as per their substance. Since website pages are something beyond plain content archives, web classification strategies need to consider utilizing other setting elements of site pages, for example, hyperlinks and HTML labels. The web is a tremendous vault of data and there is a requirement for web document classification to encourage the ordering, pursuit and recovery. Web document classification Web document classification is fundamentally not quite the same as conventional full content order as a result of the presence of some extra data given by the HTML structure. The structure-based portrayal of web documents makes utilization of only nearby data; thusly it can be utilized even in real-time classification.**

Keywords—Structured data, Unstrutured data, web document, structure vs unstructure classification, web document classification

I. INTRODUCTION

Documents classification studies are gaining more importance recently because of the availability of the increasing number of the electronic documents from a variety of sources. The resources of unstructured and semi structured information include the world wide web, governmental electronic repositories, news articles, biological databases, chat rooms, digital libraries, online forums, electronic mail and blogs repositories. So extracting information from these resources and proper categorization and knowledge discovery is an important area for research.

Web document classification has been widely studied in the past few years. Much research work has been done in this area. Chakrabarti et al. used predicted labels of neighbouring documents to reinforce classification decisions for a given document. Qi and Davison summarize the various concepts used for automatic web page classification with respect to recent works. A dynamic and various levelled arrangement framework that is equipped for including new classifications as required, sorting out the website pages into a tree structure, and characterizing pages via seeking through just a single way of the tree structure is proposed in.

The World Wide Web contains immense assets of data and administrations that keep developing quickly. Effective web crawlers have been produced to help in finding new reports by classification, contents, or subject. While it may not be currently feasible to understand the full meaning of HTML documents, intelligent software agents have been developed to extract semantic features from the words of HTML documents. These extracted features are then employed to classify and categorize the documents.

The goal of web document categorization is to classify target documents into a fixed number of predefined categories. Each document can be in multiple, exactly one, or no category at all. Using machine learning, the objective is to learn classifiers from examples that conduct the category assignments automatically, which is a supervised learning problem.
The world of computing has evolved from a small, relatively unsophisticated world in the early 1960‘s to an environment of massive size and sophistication. Everything from the daily life of individuals to our national economic productivity has been profoundly and positively affected by the growth of the use of the computer. And this growth can be measured in two ways – structured systems and unstructured systems.

Structured systems are those where the activity of processing and output is predetermined and highly organized. Structured systems are designed, built and operated by the IT department. ATM transactions, airline reservations, manufacturing inventory control systems, point of sale systems are all forms of structured systems. By contrast, unstructured systems are those that have little or no predetermined form or structure. Unstructured systems include email, reports, contracts, and other communications. A person who performs a communications activity in an unstructured system has wide latitude to structure the message in whatever form is desired. The rules of unstructured systems are fewer and less complex.

II. WEB DOCUMENT AND IT'S CLASSIFICATION

A web page, or webpage, is a document that is suitable for the World Wide Web and web browsers. A web browser displays a web page on a monitor or mobile device. The web page is what displays, but the term also refers to a computer file, usually written in HTML or comparable mark-up language. Web browsers coordinate the various web resource elements for the written web page, such as style sheets, scripts, and images, to present the web page.

Web document classification is the process of classifying documents into predefined categories based on their content. The classifiers used for this purpose should be trained from the web documents that are already classified. The task is to assign a document to one or more classes or categories. This may be done "manually" (or "intellectually") or algorithmically.

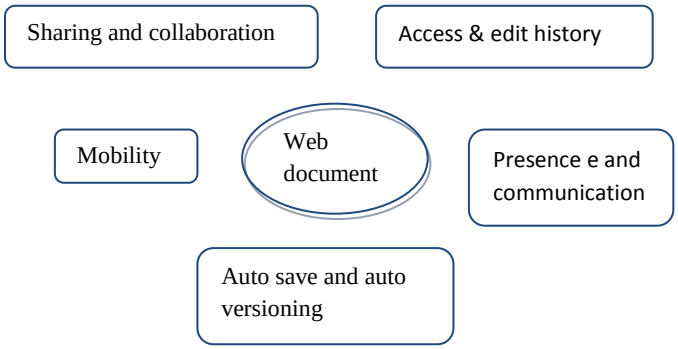


FIGURE 1: WEB DOCUMENT

A. Mobility:

This needs no clarification for web reports. Since the reports are on the web, they can be gotten to from any area/gadget.

B. Sharing & Collaboration:

This is one of the key points of interest of web documents. You can basically share your document without attaching. Along these lines, you require not stress over the software accessibility (regularly a similar rendition of software) for perusing the document at the flip side. Likewise, web docs can be cooperatively altered by various clients in the meantime without going careful numerous email strings and converging of content over different versions.

C. Presence & Communication:

At the point when a document is shared, online applications give the nearness data of the clients to whom the document is shared. On the off chance that alternate clients are on the web, you can immediately begin a chat session empowering instant communication

D. Auto-Versioning, Auto-save:

Reports develop, at the point when a document is made/altered by different clients, it experiences numerous cycles and it is constantly valuable when these documents are auto-spared and auto-versioned based on the client. You get this of course with web documents and all the more significantly a web document catches this data for all time.

E. Access & Edit History etc:

Web Documents additionally have data on when and how frequently clients got to/altered the documents and so on. This is helpful data on the off chance that you need to know whether different clients taken a look at the document you shared and so forth Manual classification cost more. The scholarly order of documents has generally been the territory of library science, while the algorithmic grouping of documents is utilized chiefly in data science and software engineering.

The issues are covering; however there is additionally interdisciplinary research on documents characterization. The documents to be grouped might be writings, pictures, music, and so on. Every sort of document has its uncommon grouping issues. Documents might be ordered by their subjects or as indicated by different traits. Web document classification is the primary requirement for search engines, which retrieve documents in response to the user query.

Documents classification or text categorization (as used in information retrieval context) is the process of assigning a document to a predefined set of categories based on the document content. Documents classification can be applied as an information filtering tool and can also be used to improve the retrieval results from a query process.

Classification is one of the main data analysis techniques and deals with the categorizing a new data entry into one of the categories based on the values of different attributes. In general, classification algorithm needs to train a model based on pre-classified training documents.

Along with search engines, topic directories (a.k.a. web directories) are the most popular sites on the Web as they are usually provided to narrow searches. Topic directories organize web pages in a hierarchical structure (taxonomy, ontology) according to their content. The purpose of this structuring is twofold. First, it helps web searches focus on the relevant collection of Web documents.

The ultimate goal here is to organize the entire web into a directory, where each web page has its place in the hierarchy and thus can be easily identified and accessed. The Open Directory Project (dmoz.org) is one of the best-known projects in this area. Second, the topic directories can be used to classify web pages or associate them with known topics. This “tagging” process can be used to extend the directories themselves.

In fact, well-known search engines such as Yahoo and Google may return with their responses the topic path, if the response URL has been associated with some topic found in a topic directory. As these topic directories are usually created manually they cannot capture all URL’s, therefore just a fraction of all responses are tagged.

III. Structure V/S Unstructured Classification

Generally, *structured information* alludes to data with a high level of association, to such an extent that incorporation in a social database is consistent and promptly accessible by basic, clear web index calculations or other inquiry operations.

A. Structured information

It alludes to sorts of information with high state of association, for example, data in a relational database. At the point when data is very organized and unsurprising, web search tools can all the more effortlessly sort out and show it in imaginative ways. Structured data markup is a content based association of information that is incorporated into a record and served from the web.

Structured information alludes to any information that lives in a settled field inside a record or document. This incorporates information contained in relational databases and spreadsheets.

B. Qualities of Structured Data:

Structured information initially relies on upon making data model – a model of the sorts of business information that will be recorded and how they will be put away, prepared and accessed. This incorporates characterizing what fields of information will be put away and how that information will be stored: data type (numeric, currency, alphabetic, name, date, address) and any limitations on the information input

Structured information has the upside of being effortlessly entered, put away, questioned and examined. At one time, as a result of the high cost and execution restrictions of capacity, memory and preparing, relational databases and spreadsheets utilizing organized information were the best way to successfully oversee information. Anything that couldn't fit into a firmly sorted out structure would need to be put away on paper in a filing cabinet.

Structured data in a DB

FIGURE 2: STRUCTURED DATA

The unstructured information is basically the inverse. The absence of structure makes assemblage a time and energy-consuming task. It is advantageous to an organization over all business strata to discover a component of data examination to diminish the costs unstructured information adds to the organization.

The expression unstructured information more often alludes to data that doesn't dwell in a customary row-column database. As you may expect, it's the inverse of organized information — the information put away in fields in a database Instances of

C. Unstructured Data:

Unstructured information documents frequently incorporate content and multimedia substance. Instances incorporate email messages, word handling records, videos, photographs, sound audio files, presentations, website pages and numerous different sorts of business archives. Take note of that while these sorts of records may have an inner structure, they are as yet thought to be "unstructured" in light of the fact that the information they contain doesn't fit conveniently in a database.

Specialists evaluate that 80 to 90 percent of the information in any association is unstructured. What's more, the measure of unstructured information in undertakings is developing altogether — regularly commonly speedier than organized databases are growing.

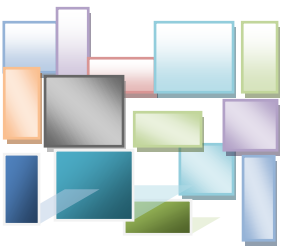
D. The Problem with Unstructured Data:

If it was conceivable or attainable to in a flash change unstructured information to structured information, at that point making insight from unstructured information would be simple. Notwithstanding, structured information is likened to machine-language, in that it makes data substantially less demanding to manage utilizing PCs; while unstructured information is (freely) more often than not for people, who don't effectively associate with data in strict, database format.

Email is an instance of unstructured information; on the grounds that while the bustling inbox of a corporate HR supervisor may be arranged by date, time or size; in the event that it were genuinely completely structured, it would likewise be orchestrated by correct subject and substance, with no deviation or spread – which is illogical, in light of the fact that individuals don't by and large talk about absolutely one subject even in focused emails.

Spreadsheets, then again, would be viewed as structured information, which can be immediately filtered for data since it is legitimately orchestrated in a relational database framework.

The issue that unstructured information presents is one of volume; most business connections are of this kind, requiring a colossal venture of assets to filter through and extricate the fundamental components, as in a web-based search engine. Since the pool of data is so vast, current data mining strategies frequently miss a generous measure of the data that is out there, quite a bit of which could be game-changing information if proficiently analyzed.



Unstructured data in a DB

FIGURE 3: UNSTRUCTURED DATA

Unstructured data represent around 80% of data. It frequently incorporates text and media content. Examples incorporate email messages, word processing reports, recordings, photographs, sound documents, presentations, website pages and numerous different sorts of business archives. Take note of that while these sorts of documents may have an inside structure, they are as yet considered « unstructured » on the grounds that the information they contain doesn't fit conveniently in a database.

Unstructured information is all over the place. Actually, most people and associations direct their lives around unstructured information. Similarly as with structured information, unstructured information is either produced by machine or created by human.

Here are few instances of machine-produced unstructured information:

- Satellite pictures: This incorporates climate information or the information that the administration catches in its satellite reconnaissance symbolism. Simply consider Google Earth, and you get the photo.
- Scientific information: This incorporates seismic symbolism, barometrical information, and high vitality material science.
- Photographs and video: This incorporates security, reconnaissance, and traffic video.

- Radar or sonar information: This incorporates vehicular, meteorological, and oceanographic seismic profiles.

Here are a few instances of human-generated unstructured data:

- **Text internal to your organization:** Think of all the content inside documents, logs, overview outputs, and e-mails. Venture data really speaks to a huge percent of the content data on the world today.
- **Social media data:** This information is produced from the web-based social networking stages, for example, YouTube, Facebook, Twitter, LinkedIn, and Flickr.
- **Mobile information:** This incorporation information such as text messages and area information.
- **Website content:** This originates from any webpage conveying unstructured content, as YouTube, Flickr, or Instagram. [2]

IV. LITERATURE REVIEW

This paper has presented a structure-based method for building high accurate web document classifier. It has demonstrated the usefulness of considering structure information, which includes

META tags, TITLE, descriptions of links and alternative texts of images. The approach is evaluated using the Bank Search dataset, and the experiments demonstrate the advantages of structure-based classification for both similar categories and distinct categories.

Compared to traditional web document classification method, combining the full text with structure information achieves nearly 6% accuracy improvement in the case of similar categories and 3.7% accuracy improvement in the case of distinct categories. Results also demonstrate that support vector machine is very suitable for web document classification.

We attempted automatic classification of unstructured blog posts using a semi-supervised machine learning approach. Empirical studies indicate that the multi-step classification strategy outlined can classify blog text with good accuracy. We confirmed that the combination of tf-idf and multi-word heuristics is an effective statistical feature-set extractor for blog entries.

Moreover, our empirical results indicate that the naïve Bayesian classification model clearly out-performs the basic ANN based classification model for highly domain-dependent unstructured blog text classification especially when a restricted feature-set is available.

However, we would like to repeat our experiments with larger and more varied datasets. We would also like to investigate the effect of changing neural network configuration on blog text classification accuracy.

The growing phenomenon of the textual data needs text mining, machine learning and natural language processing techniques and methodologies to organize and extract pattern and knowledge from the documents. This overview focused on the existing literature and explored the automatic documents classification documents representation and knowledge extraction techniques. Text representation is a crucial issue. Most of the literature gives the statistical of syntactic solution for the text representation.

However the representation model depend on the informational that we require. Concept base or semantically representations of documents require more research. Several algorithms or combination of algorithms as hybrid approaches are proposed for the automatics classification of documents, among these SVM and NB classifier are shown most appropriate in the existing literature. However more research is required for the performance improvement and accuracy of the documents classification and new method to solutions are required for useful knowledge from the increasing volume of electronics documents.

The following are the some opportunities of the unstructured data classification and knowledge management:

- To reduce the training and testing time and improve the classification accuracy, precision, recall, micro-average macro-average.
- Spam filleting and email categorization: User may have folders like, electronic bills, email from family and friends, and so on, and may want a classifier to classify each incoming email and automatically move it to the appropriate folder. It is easier to find messages in sorted folders than in a very large inbox.
- Automatic allocation of folders to the downloaded articles, documents from text editors and from grid network.
- Semantic and Ontology: The use of semantics and ontology for the documents classification and informational retrieval.
- Trend mining i.e. marketing, business, and financial trend (stock exchange trend) form e-documents (Online news, stories, views and events).
- Mining text streams: Some new techniques and methods are required for handling stream text.

In this study, Naïve Bayes classifier has been discussed as the best document classifier, which satisfies the literature result. Through the implementation of different feature selection and classifier available in WEKA, it is demonstrated preprocessing and feature selection are two important steps to improve the mining quality.

There are many words in the documents, therefore when we captured the terms from these documents, thousands of terms are found. However, there are some terms that are usefulness and uninteresting to the results, it is then important to discover and interpret which features are useful and critical. Concerning numerous searching and selection techniques are available; it is encouraged to apply all these techniques and hence selects the best one for preprocess the data as well as to build the model.

Furthermore, the performance of mining result is directly affected by the quality of data. So, preprocessing phase is important to make the data being more precise (so as to achieve a better classification result) and even improve the time used to train and general the model, as proven in the experiment section

V. CONCLUSION

In this paper we briefly describe the structured and unstructured data. Also this paper concisely explains the structured vs. unstructured classification, the problem which occurs in unstructured data, types of unstructured data. Also we outline the web document classification and characteristics of web document.

REFERENCES

[1] Rajendra Kumar Roil, S.K SahayBITS, Pilani - K.K. Birla, Goa Campus Zuarinagar, Goa - 403726, India, "An Effective Approach for Web Document Classification using the Concept of Association Analysis of Data Mining".

[2] Aurangzeb Khan , Baharum B. Bahurdin, Khairullah Khan Department of Computer & Information Science Universiti Teknologi, PETRONAS, "An Overview of E-Documents Classification", 2009 International Conference on Machine Learning and Computing IPCSIT vol.3 (2011) © (2011) IACSIT Press, Singapore

[3] Mita K. Dalal , Mukesh A. Zaveri, Sarvajanik College of Engineering & Technology, S. V. National Institute of Technology, Surat, India. "Automatic Classification of Unstructured Blog Text", Journal of Intelligent Learning Systems and Applications, 2013

[4] Kejing He, Chenyang Li,School of Computer Science and Engineering, South China University of Technology, Guangzhou , China, "STRUCTURE-BASED CLASSIFICATION OF WEB DOCUMENTS USING SUPPORT VECTOR MACHINE", Proceedings of CCIS2016,IEEE

[5] S.L. Ting, W.H. Ip, Albert H.C. Tsang Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hung Hum, Kowloon, Hong Kong, "Is Naïve Bayes a Good Classifier for Document Classification?", International Journal of Software Engineering and Its Applications Vol. 5, No. 3, July, 2011

[6] S. Chakrabarti, B. Dom, and P. Indyk. Enhanced hypertext categorization using hyperlinks. In SIGMOD '98: proceedings of the 1998 ACM SIGMOD International Conference on Management of Data, pages 307–318, New York, NY, USA, 1998.

[7] Qi, X. and B. D. Davison (2009). "Web page classification: Features and algorithms." ACM Computing Surveys (CSUR) 41(2): Article No.: 12.

[8] Xiaogang Peng, Ben Choi (2002), "Automatic Web Page Classification in a Dynamic and Hierarchical Way", In proceedings of IEEE, Second International Conference on Data Mining, Washington DC, IEEE Computer Society, pp: 386-393.

[9] Bright Planet, "Structured Vs Unstructured Data" [Online] Available: https://brightplanet.com/2012/06/structured-vs-unstructured-data/

[10] J. Ronk, "Structured, Semi structured and Unstructured data" [Online] Available: https://jeremyronk.wordpress.com/2014/09/01/structured-semi-structured-and-unstructured-data/

[11] Google Search, "Introduction to structured data" [Online] Available:https://developers.google.com/search/docs/guides/intro-structured-data

[12] Vangie Beal, "Structured data" [Online] Available: http://www.webopedia.com/TERM/S/structured_data.html

[13] B. Inmon, "Unlocking the potential of unstructured content" [Online]. Available: http://inmoncif.com/registration/whitepapers/unlocking-final.pdf

[14] Wikipedia, "Web page" [Online]. Available: https://en.wikipedia.org/wiki/Web_page

[15] Raju Vegesna, "Document vs Web Document" [Online]. Available: https://www.zoho.com/general/blog/document-vs-web-document.html