INFORMATION AGGREGATION FROM VARIOUS WEBSITES USING WEB CONTENT MINING TECHNIQUES: A SURVEY

NARENDRA M. JATHE¹, Dr. AVINASH B. MANWAR², Dr. HEMANT S. MAHALLE³

1. Department of Computer Science, Arts, Commerce and Science College, KiranNagar, Amravati.
2. Department of Computer Science, P. G. T. D, Sant Gadge Baba, University, Amravati.

Accepted Date: 05/03/2015; Published Date: 01/05/2015

Abstract: User tends to search the information for a particular domain and as although the desired information is available on Web; it is in very scattered and varied structures across different geographic locations. Aggregating information in such a complex structure of Web is a challenging task. The combination of characteristics such as comparison, relation, intra-organization and inter-organization aggregation to maximize values in the aggregated information is really a challenging task. Although wrapping technology such as screen scraping, direct data feed with some encoding standard such as XML have been used by research groups in this field, there is a wide scope for research in information aggregation to facilitate varied aggregating services. This paper presents a survey and emerging techniques on information aggregation from various websites using web content mining techniques. Aggregation of information plays a vital role in the construction of knowledge based systems in various domains, ranging from healthcare, economics, applied science, artificial intelligence, and robotics to decision-making processes and machine learning.

Keywords: Web Content Mining, Information Extraction, Information aggregation, web document types, Mining techniques and Attribute Extraction.

Corresponding Author: MR. NARENDRA M. JATHE

Access Online On:
www.ijpret.com

How to Cite This Article:
Narendra M. Jathe, IJPRET, 2015; Volume 3 (9): 865-872

Available Online at www.ijpret.com
INTRODUCTION

The amount of information from conventional databases and web sources has been growing exponentially in the past few years. Effective use of this huge amount of information has become a major challenge to the research communities working in this area. As far as the frequent changes in technologies and services are concerned, existing business concerns are tending towards making adaptive transformations so as to aggregate and integrate values from relevant information sources. Information aggregation is a service that gathers relevant information from multiple sources to provide convenience and add value by analyzing the aggregated information for specific objectives using Internet technologies. The providers of this service are called “ aggregators” in general. In a broader sense, information intermediaries such as newspapers, magazines, professional journals, and more recently, increasing number of web portals are information aggregators since they all collect information from multiple sources and disseminate it for convenient consumption [1].

Universities have made their public information available through their established web sites. As far as the contents of this particular domain is considered, most of the parametric quantity of information is common in nature and visitors of these site are ranging from students, academicians, researchers to common people. But the information across these sites is very scattered and having wide scope to aggregate information in this domain and a very few research communities are working towards this domain for information aggregation.

II. EMERGING TECHNIQUES OF INFORMATION AGGREGATION FROM VARIOUS WEBSITES USING WEB CONTENT MINING TECHNIQUES

Information aggregation, a service that collects relevant information from multiple sources, has emerged to help individuals and businesses to effectively use the growing amount of information on the Web. It analyzed a number of characteristics of information aggregation, namely comparison, relationship, and intra-organization aggregation.

Bemadette Bouchon-Meunier (1998) presented the tools for information aggregation considering multiple criteria, fusion of data provided by several sources, imperfect knowledge availability. Author includes theoretical and applied studies of aggregation and fusion methods in the frameworks[2].

Michael W. Berry (2003) introduced the approaches to capture the semantics of text-based document collections based on Bayesian models, probability theory, vector space models, and statistical models[3].
Kokkinos, P. (2009) proposed information aggregation as a method for summarizing the resource-related information, used by the task scheduler. In this paper variety of techniques are described for aggregating the information of the resources belonging to a hierarchical Grid domain. Authors have claimed that the proposed aggregation schemes achieved large information reduction, while enabling good task scheduling decisions by means of simulation experiments[4].

Manzoor, A. (2009) presented a context aggregation system that detects and removes the duplicates and conflicts from context information by using the policies based on quality of context parameters. Authors have claimed that this system effectively aggregates the continuously evolving context information and efficiently uses the scare resources in pervasive environments[5].

The intuition of data reliability has been incorporated into the mainstream of research on ordered weighted averaging (OWA) operators. Instead of relying on human-guided variables, the aggregation behavior is determined in accordance with the underlying characteristics of the data being aggregated. Data-oriented operators such as the dependent OWA (DOWA) utilize centralized data structures to generate reliable weights. Despite their simplicity, the approach taken by these operators neglects entirely any local data structure that represents a strong agreement or consensus. To address this issue Boongoen, T. (2010) proposed the cluster-based OWA (Clus-DOWA) operator. It employs a cluster-based reliability measure that is effective to differentiate the accountability of different input arguments. This paper presented a nearest-neighbor-based reliability assessment avoiding an expensive clustering process. The proposed measure can be perceived as a stress function, from which the OWA weights and associated decision-support explanations can be generated. To illustrate the potential of this measure, authors have applied it to both the problem of information aggregation for alias detection and the problem of unsupervised feature selection[6].

Famaey, J. (2011) proposed a context dissemination framework. The framework facilitates the aggregation and translation of information as it is propagated through the hierarchy. Additionally, by way of semantics, context is filtered based on meaning and is disseminated intelligently according to dynamically changing context requirements. This significantly reduces the exchange of superfluous context and thus further increases scalability. The authors have claimed that the large size of modern federated cloud computing infrastructures makes the presented context dissemination framework ideally suited to improve their management efficiency and scalability. An evaluation of the framework have been claimed by the authors to characterize the benefits of given approach in terms of scalability and reasoning time[7].
Ishii, H. (2012) have developed a method to aggregate the web pages into groups by exploiting the sparsity inherent in the web. The method is motivated by results on singular perturbation techniques for large-scale Markov chains and multi-agent consensus[8].

Gupta, R. (2012) discussed techniques for executing aggregation queries over distributed data to minimize the number of message exchanges between data sources, aggregators, and users. Authors examined the problem in terms of different types of queries, aggregation functions, query imprecisions, and whether the aggregators get data from sources using pull- or push-based mechanisms[9].

The disparate communities of climate modeling and remote sensing are finding economic, political, and societal benefit from the direct comparisons of climate model outputs to satellite observations. In this context, Crichton, D.J. (2012) studied principals to build a software infrastructure that enables these comparisons. The authors have focused on the description of software tools and services that meet the challenges of navigating and aggregating information from the largely distributed ecosystem of organizations[10].

Wei, Chunfu(2012) suggested a methodology for solving multiple attribute decision making problems in which the decision information is the fuzzy linguistic value. The methodology is based on the EWOA and LHA operators supported by an illustrative example[11].

Electronic Health Records (EHR) applications have evolved a lot in recent years taking advantage of the strong growth of information and communication technologies (ICT) in the healthcare sector. With the existence of multiple solutions to solve the problems of systemic interoperability would be expected a greater and quicker rise of these applications within, and beyond, healthcare systems. Carvalho, S. (2012) presented a platform to support applications in the social sector, with the intention of guarantee the interoperability of them with the existing healthcare information systems. This platform is likely to generate new information objects to accommodate the needs of recording and sharing of information from a wide range of providers[12].

El-Tawab, S. (2012) has introduced the theoretical foundations of FRIEND: A cyber-physical system for traffic Flow-Related Information aggrEgatioN and Dissemination. Authors have claimed that by integrating resources and capabilities at the nexus between the cyber and physical worlds, FRIEND will contribute to aggregating traffic flow data collected by the huge fleet of vehicles on roads into a comprehensive, near real-time synopsis of traffic flow conditions. Authors have envisioned the cat-eye system to be supplemented by road-side units (RSU) deployed at regular intervals. The RSUs placed on opposite sides of the roadway
constitute a logical unit and are connected. The physical components of FRIEND collect traffic flow-related data from passing vehicles. The collected data is used by an inference engine in the RSU's cyber component to build beliefs about the state of the traffic, to detect traffic trends, and to disseminate relevant traffic flow-related information along the roadway[13].

Aggregation of fuzzy information is a new branch of Atanassov's intuitionistic fuzzy set (AIFS) theory, which has attracted significant interest from researchers in recent years. In this paper Weize Wang (2012) has treated the intuitionistic fuzzy aggregation operators with the help of Einstein operations. Some new operations of AIFSs, such as Einstein sum, Einstein product, and Einstein scalar multiplication were introduced. Authors have developed some intuitionistic fuzzy aggregation operators, such as the intuitionistic fuzzy Einstein weighted averaging operator and the intuitionistic fuzzy Einstein ordered weighted averaging operator, which extend the weighted averaging operator and the ordered weighted averaging operator to aggregate Atanassov's intuitionistic fuzzy values, respectively. Numerical examples were given to illustrate the developed aggregation operators[14].

Tasaki, Y. (2012) described an information aggregation system that captures partial contents of a Web page across multiple Web pages. The proposed system called Aggnel is implemented as a prototype system which aggregates partial contents from multiple Web pages. The proposed method is evaluated by means of an experiment allowing users to aggregate key information[15].

Meimei Xia (2013) studied the aggregation of intuitionistic multiplicative preference information, proposed some aggregation techniques, investigated their properties, and applied them to decision making based on intuitionistic multiplicative preference relations[16].

Casado, A.M. (2013) presented an alternative decision model where the information fusion processes are a function of the aggregate values in order to model the perception group and aggregate the expert’s opinion in corporate reputation evaluation problems[17].

Xindong Wu (2014) presented a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. Authors have analyzed the challenging issues in the data-driven model and also in the Big Data revolution[18].
III. CONCLUSIONS

In this paper, we have surveyed large number of literature from journal papers, magazine articles, white papers, etc. starting from year 1998 to 2014. In section II of this paper, we have categorized the research work of information aggregation carried out by these research communities and quoted representative work in accordance with the categories. Many parameters such as document semantics, imperfect knowledge availability, data fusion across disparate platforms and hierarchy, changing context, inherent sparsity of web, interoperability have been considered by the researchers and due to very scattered nature of web there is still wide scope of augmenting these parameters for improvisation. Theoretical and applied approaches taken by the researchers with tools and specific systems, prototypes such as FRIEND, Aggnel in context to information aggregation have been surveyed in this paper.

Although this paper surveyed varied parameters, methodologies, case studies, it is very apparent conclusion that information aggregation must dealt with the specific domain considering relative parameters and nature of data in addition to universal parameters discussed.

REFERENCES


18. Xindong Wu; Xingquan Zhu; Gong-Qing Wu; Wei Ding. “Data Mining with Big Data”. IEEE Transactions on Knowledge and Data Engineering, Vol. 26, Issue 1, pages 97–107, 2014.