Text Mining-Based Semantic Web Architecture (TM-SWA) for e-Learning Systems

Hamad Ibrahim Alomran

***Abstract*—This paper highlights semantic web techniques and proposes architecture for e-Learning-based systems for the academic portal. Text mining is used with the proposed model for better processing of unstructured data available in XML and RDF formats. An algorithm will be used to support building a web retrieval system to extract the hidden knowledge for the semantic web by ontologies for e-learning items to classify and find the relationships between the leaning items via the academic portal.**

***Index Terms*—Semantic web, text mining, data mining, e-learning, web, web architecture.**

I. Introduction

Databases available on the web allow search capability in the information provided in these databases and its retrieval; however, they do not build new knowledge derived from this information. Therefore, these databases do not offer the possibility of linking the information and the discovery of new knowledge. Many educational institutions maintain e-learning portals and generate gigabytes of data every year. This is the appropriate time for these institutions to use Semantic Web and improve the web content. Despite the fact that the web is rich with information, gathering and analysing data to obtain useful result is a difficult task. Documents on the web are largely unorganised. In the coming decades, the biggest challenge will be how to effectively and efficiently dig out a machine-understandable and readable information and knowledge layer, called the Semantic Web, from unorganised, human-readable web data. This problem has motivated specialists to search for new methods and techniques for information retrieval, and also thinking of a new approach for solving this issue. The target is not retrieving available information on the web, but the creation of information and new knowledge using available information and data, and this is done through linkage between information [1].

The aim of this paper is to build an information system on the web to extract knowledge from the available information through the designed portal. The result helps the user to obtain the greatest amount of possible integrated and interrelated information. In this paper, a framework will be introduced to build a web retrieval system for extracting hidden knowledge from semantic web. This contributes to databases' knowledge consisting Semantic Web, find the relationships between them, and create knowledge through data (web) mining in Semantic Web technology.

This paper includes two important sections. First, this is one of few studies that deal with web mining in the Semantic Web. It is expected that the result of this study enriches cognitive stock in its field and opens the way for future studies dealing with other aspects not addressed in this study. The second side lies in building an information system available on the Web where information and knowledge can be retrieved as well as to discover new knowledge through data mining technology in Semantic Web. In this regard, building information systems on the web permits entering bibliographic data to include the name of the author, title of the study, periodical name, number, month, year, page numbers and keywords.

II. Literature Review

In the literature published about this topic, most studies address the issue of data mining, and the semantic web, whereas few others studies deal with the issues that link data mining within the Semantic Web.

One of the important studies in this regard is conducted Rodr ǵuez *et al.* under the title of 'Ontologies-driven web mining: concepts and techniques'. The study points out that the aim of semantic web mining is to combine the development of two research areas, namely semantic web and web mining [2].

Dhenakaran and Yasodha's [3] 'Semantic Web Mining—A Critical Review' highlights the aims of the Semantic Web to address the problem of web retrieval systems (WIR) by providing machine interpretable semantics to provide greater machine support for the user. These two areas pave way for the extraction of relevant and meaningful information from the web, thereby giving rise to the term 'Semantic Web Mining'.

Stummea *et al.* [4] explain the convergence of trends from two aspects: improving the results of web mining by exploiting semantic structures in the web, and using web-mining techniques to build the Semantic Web.

In a study by Aufaure *et al.* [5] entitled 'Metadata- and Ontology-based Semantic Web Mining', the authors state that the increasing volume of data available on the web makes information retrieval a tedious and difficult task. The vision of the Semantic Web introduces the next generation of the web by establishing a layer of machine-understandable data (e.g. for software agents, sophisticated search engines and web services). To this end, the success of the Semantic Web depends on the easy creation, integration and use of semantic data.

Sampson *et al.* [6] presented a study under the title of

Manuscript received January 2, 2014; revised May 7, 2014.

Hamad Ibrahim Alomran is with the Department of Information Management College of Computer and Information Sciences Al-Imam Muhammad Ibn Saud Islamic University, Riyadh (e-mail: alomran@imamu.edu.sa).

'Ontologies and the Semantic Web for E-learning'. They point out that the Semantic Web is the emerging landscape of new web technologies aimed at web-based information and services that would be understandable and reusable by both humans and machines. Ontologies are generally defined as a representation of a shared conceptualisation of a particular domain that represents a major component of the Semantic Web. The study shows that ontologies and Semantic Web technologies will influence the next generation of e-learning systems and applications.

Data integration applications offer the potential to connect disparate sources, but they require one-to-one mappings between elements in each different data repository. However, the Semantic Web allows a machine to connect to any other machine and efficiently exchange and process data based on built-in, universally available semantic information that describes each resource. In effect, the Semantic Web will allow users to access all of the information listed above as one large database [7].

It is important to point out that the concept of ontology is central to the Semantic Web project. Web pages are considered to comprise statements that relate objects. The denotations of the termscomposing the statements need to be fixed relative to a particular universe of discourse represented in ontology, which codifies a shared and common understanding of some domain [8].

Ypma and Heskes [9] propose a method for learning content categories from usage. They model navigation in terms of hidden Markov models, with the hidden states being page categories and the observed request events being instances of them. Their primary aim is to show that a meaningful page categorisation may be learned simultaneously with theuser-labelling and inter-category transitions; semantic labels (such as 'sports pages') must be assigned to a state manually. The resulting taxonomy and page classification can be used as a conceptual model for the site or to improve an existing conceptual model.

Another application uses mining techniques to learn a classification of site users' goals from their navigation patterns. Ed Chi *et al*. identify frequent paths through a site. Based on the keywords extracted from the pages along the path, they compute the likely information sent followed (i.e. the intended goal of the path). The information that is sent is a set of weighted keywords that can be inspected and labelled more concisely using an interactive tool. Thus, usage creates a set of information goals that users expect the site to satisfy [10], [11].

Most studies have addressed the issues of data mining and the Semantic Web, but few have dealt with the two issues linked as data mining in the Semantic Web.

One of the most important studies in the field of data mining was conducted under the title of 'Semantic Web Mining', in which the authors state that Semantic Web mining aims to combine the development of two research areas: Semantic Web and Web mining and using this combination to represent and apply the result of the algorithm to great the semantic items [3].

Zhou, Hui and Fong, in a paper entitled 'Web Usage Mining for Semantic Web Personalization', state that it has become more difficult to access relevant information from the Web due to the explosive growth of information. One possible approach to solve this problem is web personalisation. In the Semantic Web, user access behaviour models can be shared as ontology. Agent software can then utilise it to provide personalised services such as recommendations and searches [10].

In 'Mining the Semantic Web', Chakravarthy presents research on how Semantic Web technologies can be used to mine the web for information extraction. He also examines how new, unsupervised processes can aid in extracting precise and useful information from semantic data, thus reducing the problem of information overload. The researcher points out that the Semantic Web adds structure to the meaningful content of web pages; hence, information is given a well-defined meaning that is both readable by humans and can be processed by machines [12].

III. Semantic Web

The vision of the Semantic Web is a 'web of data' that not only harnesses the seemingly endless amount of data on the World Wide Web, but that also connects information with data in relational databases and other non-interoperable information repositories(i.e. EDI systems). Considering that relational databases now house the majority of enterprise data, the ability of Semantic Web technologies to access and process these data alongside other data from web sites, other databases, extensible markup language (XML) documents and other systems exponentially increases the amount of useful data available. Databases are organised in tables and columns based on the relationships between the data they house, and these relationships reveal the meaning (semantics) of the data [13].

Fig. 1 represents the architecture of a sample Semantic Web application, which is used in this paper to develop a model. As the figure illustrates, the architecture comprises the following components:

- resource description framework (RDF) triple store,
- dynamic content engine,
- artificial intelligence (AI) application,
- browser [14].

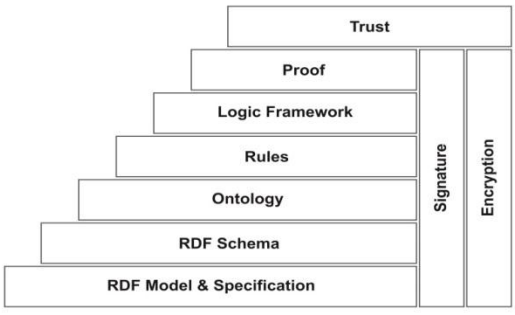

Fig. 1. Architecture of a sample of semantic web application.

The real power of the Semantic Web will be realised when people create many systems that collect web content agents. Thereby, the effectiveness of the Semantic Web will increase drastically as more machine-readable web content and automated services become available. This level of inter-agent communication will require the exchange of proofs. There are two important available technologies

considered in this paper for developing the semantic web:XML and RDF.

XML allows users to create their own tags to annotate web pages or sections of texts on a page. Programs can use these tags in sophisticated ways, but the programmer must know what the page writer uses each tag for. XML permits users to add arbitrary structure to their documents, but it says nothing about what the structures mean. The meaning of XML documents is intuitively clear because of 'semantic' markup and tags, which are domain-terms [15].

RDF provides a means for adding semantics to a document. RDF is an infrastructure that enables the encoding, exchange and reuse of structured metadata. A process in which semantic information is added to web documents is called semantic annotation. In combination with RDFS, RDF offers modelling primitives that can be extended according to the needs at hand [14].

A. Ontology Supports for the Semantic Web

Ontology is an explicit specialisation of a conceptualisation. Ontology's specifications are conceptualisation and corresponding vocabulary used to describe a domain. They are well suited for describing heterogeneous, distributed and semi-structured information sources that can be found on the web. By defining shared and common domain theories, ontology helps both people and machines to communicate concisely by supporting the exchange of semantics and not only syntax. This is important because semantics for the web is based on an explicitly specified ontology [16].

B. Modelling the Semantics of the Web

According to the W3C recommendation, the RDF 'is a foundation for processing metadata; it provides interoperability between applications that exchange machine-understandable information on the web'. RDF documents consist of three types of entities: resources, properties and statements. Resources may be web pages, parts or collections of web pages, or any (real-world) objects that are not directly part of the World Wide Web. Resources are always addressed by URIs. Properties are specific attributes, characteristics or relations that describe resources. A resource together with a property that has a value for that resource forms an RDF statement.

There are four principles which must be taken into account during developing a semantic web application:

- All the data and entries that share the same information should be identified by Uniform Resource Identifier (URI) references.
- The data must be provided in RDF format.
- The URI in Hypertext Transfer Protocol (HTTP) should be linked to the RDF that belongs to it.
- The data should be interlinked with each other.

Architecture of a sample of semantic web application (see Fig. 1) has the following components:

- RDF triple store.
- Dynamic content engine.
- Artificial Intelligence (AI) application.
- Browser.

According to the Fig. 2, the middle part shows an example of RDF statements. The data model underlying RDF is basically a directed labelled graph. RDF schema defines a simple modelling language on top of RDF that includes classes, isa relationship between classes and between properties, and domain/range restrictions for properties. RDF and RDF schema are noted with the syntax of XML, but they do not employ the tree semantics of XM [17].

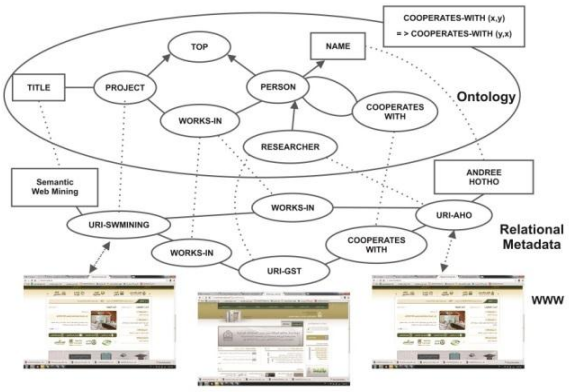

Fig. 2. The relation between the World Wide Web, relational metadata, and ontologies.

C. Extracting Semantics from Web Usage

The Semantic Web adds semantic annotations to web documents in order to access knowledge instead of unstructured material. This allows knowledge to be managed in an automatic way. Web mining helps to learn definitions of structures for knowledge organisation, such as ontologies, and it also provides the population of such knowledge structures.

The combination of implicit user input (usage) and explicit user input (search engine queries) can serve as a further generator of conceptual structure. User navigation has been employed to infer topical relatedness (i.e. the relatedness of a set of pages to a topic as given by the terms of a query to a search engine). Aggarwal [18] refers to this as 'collaborative crawling' and proposes it as a method for improving on focused and intelligent crawling, which uses only the information from page content and hyperlink structures. Classifying pages into 'satisfying the user defined predicate' and 'not satisfying the predicate' is thus learned from usage, structure and content information. An obvious application is to mine user navigation to improve search engine rankings [19].

IV. Semantic Web for e-Learning Systems

E-learning covers a wide set of applications and processes, including web-based learning, computer-based learning, virtual classrooms and digital collaboration. It includes the delivery of content via Internet and intranet/extranet (LAN/WAN), audio and videotape, satellite broadcast interactive TV and CD-ROM. E-Learning provides faster learning at reduced costs, increases access to learning and provides clear accountability for all participants in the learning process.

- Effective e-learning is available on demand (i.e. 24x7 availability anytime, anywhere to anyone). This brings

about some characteristics that in turn provide benefits to users.

- E-learning allows users to take ownership of their own training and development and provides HUDwith personalised learning/training roadmaps and easy access to applications and approvals online.
- E-learning allows users to connect and obtain information globally.
- E-learning encourages users to obtain learning continuously, as opposed to the traditional educationlifestyle, thus promoting the concept of life-long learning.
- Users obtain a total learning experience that is highly interactive and 'hands-on',directed by their own content and time.

TABLE I: BENEFITS OF USING SEMANTIC WEB TECHNOLOGY FOR E-LEARNING [23]

Characteristics	E-learning	Semantic web
Delivery	Pull: student determines Agenda	Knowledge items (learning materials) are distributed on the web, but they are linked to commonly agreed ontologies, which enables the construction of a user-specific course via semantic querying for topics of interest.
Responsiveness	Reactionary: respondsto problem at hand	Software agents on the Semantic Web may use a commonly agreed service language that enables coordination between agents and the proactive delivery of learning materials in the context of actual problems. The vision is that each user has his or her own personalisedagent that communicates with other agents.
Access	Non-linear: allowsdirect accesstoknowledgein whatever sequence makes sense to the situation	Users can describe the situation at hand and perform semantic querying for the suitable learning material. The user profile is also accounted for, and access to knowledge can be expanded by semantically defined navigation. An example student, teacher .
Symmetry	Symmetric:learning occurs as an integrated activity	The Semantic Web (semantic intranet) offers the potential to become an integration platform for all business processes in an organisation, including learning activities.
Modality	Continuous:learning runs parallel to business tasks and never stops	Active delivery of information (based on personalisedagent) creates dynamic learning environments that are integrated in the business processes.
Authority	Distributed:content comes frominteractionof participants and educators	The Semantic Web will be as decentralisedas possible. This enables effective cooperative content management.
Personalisation	Personalised content is determined by individual users' needs and aims to satisfy all users' needs.	A user (using his or her personalisedagent) searchesforlearningmaterial that is customised to his or her needs.The ontology is the link between users' needs and characteristics of the learning material.
Adaptively	Dynamic: content changes constantly through user input, experiences,newpractices, business rules and heuristics.	The Semantic Web enables the use of distributed knowledge provided in various forms, enabled by semantic annotations of content. The distributed nature of the Semantic Web enables the continuous improvement of learning materials.

Using e-learning methods for training and development allows individuals to define and populate metrics for successful learning (upon completion) and to link these metrics to their performance in an organisation [20], [21].

Table I shows the comparison between a learning and semantic learning based on the main elements of learning process.

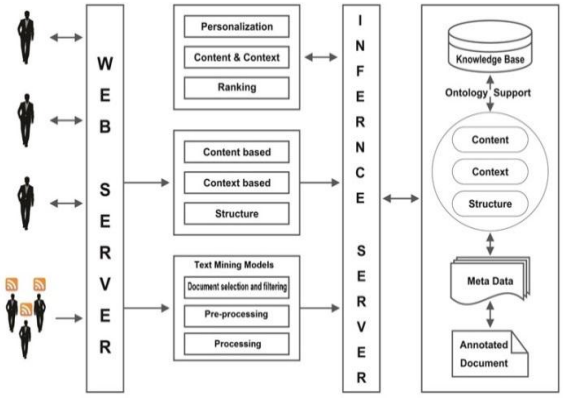

Fig. 3. Scenario of semantic web-based e-learning.

V. ARCHITECTURE FOR SEMANTIC WEB BASED E-LEARNING: AN EXAMPLE

Based on the discussion in the previous section, this section presents the overall architecture of our ontology-based e-learning scenario. The architecture of the system is presented in Fig. 3. The knowledge warehouse acts as a metadata repository, and the OntoBroker system is a principal inference mechanism. The primary activities in an e-learning environment are:

- providing information from authors
- accessing learning materials by readers and authors by querying and browsing

The ontology and semantics are implemented in the initial steps. The inference server also maintains a very good knowledge base, which is referred to as the archive. This archive is used as the main input for the text mining. The data generated by the e-learning systems are periodicallymonitored by the inference servers. The text mining models are available as part of architecture processes all of the available data in the form of XML or RDF documents, and they present the data to the web servers. The contents that are processed by the text mining models are personalisedand ranked based on the available historical data.

A. Providing the Learning Materials

The first phase is the production of learning materials that may be used or reused in the construction of training courses. Toprovide learning materials that willbesuitablefor metadatasearching, each learning material must be described or 'enriched' with the following metadata information:

- What is the learning material about (content annotations)?
- Which is the context of the learning material (context annotations)?
- How is it connected to other learning materials (structure annotations)?

B. Accessing the Learning Materials

Intheprocessofaccessing information, the ontology is

used for:

- Semantic querying for learning materials is based on the three-dimensional search spaces, content, context and structure, which are defined by the ontology. An easy-to-use interface based on the query capabilities of the F-logic query interface of OntoBroker is offered for specifying such queries.
- Conceptual navigation through the collection of learning materials based on ontological relations between concepts in the:
  a) content,
  b) context ontology's.

VI. Role of Text Mining in the Developed Framework

Text mining is the analysis of data contained in natural language text. The application of text mining techniques to solve business problems is called text analytics. Text mining can help organisations derive potentially valuable business insights from text-based content such as Word documents, emails and postings on social media streams such as Facebook, Twitter and LinkedIn. However, mining unstructured data with natural language processing (NLP), statistical-modelling and machine-learning techniques can be challenging because natural language text is often inconsistent [22].

A. Text Mining Stages Used in the Developed Architecture

- Document selection and filtering (IR techniques),
- Document pre-processing (NLP techniques),
- Document processing (NLP/ML/statistical techniques).

Document selection involves the identification and retrieval of potentially relevant documents from a large set (e.g. the web) in order to reduce the search space. Standard or semantically enhanced IR techniques can be used for this. Document pre-processing involves cleaning and preparing the documents (e.g. removal of extraneous information, error correction, spelling normalisation, tokenisation, POS tagging). Document processing mainly consists of information extraction. For the Semantic Web, this is realised in terms of metadata extraction [24].

B. Metadata Extraction

There are two types of metadata extraction. Explicit extraction involves information that describes the document, such as that contained in the header information of HTML documents (e.g. titles, abstracts, authors, creation date). Implicit extraction involves semantic information that is deduced from the material itself (i.e. endogenous information such as names of entities and relations contained in the text). This essentially involves information extraction techniques, often with the help of ontology [23].

Fig. 4 and Fig. 5 show screenshots of the e-learning management system (Desire2Learn) implemented at Majmaah University, Kingdom of Saudi Arabia. Indeed, the figures are quite self-explanatory.

The current system (D2L) controls students and instructors enrolment. However, instructors need to add course materials themselves. The proposed framework present a system utilize ontology and builds clusters based on student interests. Those clusters enable students to collaborate, exchange information, and share knowledge via virtual discussion rooms and other web-based tools. Finally creating ontology stored in the knowledge base,then build semantic web for e-learning process.

Fig. 4. E-learning portal.

Fig. 5. E-learning processes.

VII. Conclusion

This paper discusses new Semantic Web architecture. The Semantic Web is a relatively new sub-field of Web 3. It has a vast scope for investigation considering the availability of a large amount of unstructured data on the web. The non-availability of a rugged database management system to manage the Semantic Web opens up new avenues for those in the field of study to develop a Knowledge Extraction Management System (KEMS) for unstructured data available on the web. In this paper, we have developed architecture for Semantic Web mining with the help of text mining tools. The paper illustrates how text mining can improve the results of the underlying architecture by exploiting the new semantic structures in the web. Further, it illustrates how the construction of the Semantic Web can use text mining techniques.

References

[1] W3C site. [Online]. Available: http://www.w3c.org
[2] R. González, A. G. Crespo, R. C. Palacios, and J. M. G. Berbis, "Using caching techniques to improve the performance of inference applications in semantic technologies," in *Ontologies-driven web mining: Concepts and techniques*, H. Oscar Nigro and S. Gonzalez-Cśaro Eds., 2011.
[3] S. S. Dhenakaran and S. Yasodha, "Semantic web mining: A critical review," *International Journal of Computer Science and Information Technologies*, 2011, vol. 2, no. 5, pp. 2258–2261.
[4] G. Stummea, A. Hotho, and B. Berendt, "Semantic web mining," State Of The Art And Future Directions A Knowledge And Data Engineering Group, University of Kassel, Institute of Information Systems, Humboldt University, Berlin, 2006.
[5] M. A. Aufaure, B. L. Grand, M. Soto, and N. Bennacer, "Metadata- and ontology-based semantic web mining," in *Web semantics & ontology*, D. Taniar and J. W. Rahayu, Eds., 2006, pp. 259–296.

[6] G. Sampson, M. D. Lytras, G. Wagner, and P. Diaz, "Ontologies and the semantic web for e-learning," *Educational Technology & Society*, vol. 7, no. 4, pp. 26–28.
[7] Altova. (2011). What is semantic web? [Online]. http://www.altova.com/semantic_web.html
[8] P. Clerkin, P. Cunningham, and C. Hayes, "Ontology discovery for the semantic web using hierarchical clustering," in *Proc. The Semantic Web Mining Workshop Co-Located With ECML/PKDD*, 2001.
[9] A. Ypma, J. Geurts, S. Özer, E. van deer Werf, and B. de Vries, "Online personalization of hearing instruments," *Eurasip J. Audio, Speech and Music Processing*, 2008.
[10] B. Zhou, S. C. Hui, and A. C. M. Fong, "A web usage lattice based mining approach for intelligent web perzonalization," *International Journal of Web Information Systems*, vol. 1, no. 3, pp.137–146, 2005.
[11] H. Chi, P. Pirolli, K. Chen, and J. Pitkow, "Using information scent to model user information needs and actions on the web," in *Proc. the ACM Conference on Human Factors in Computing Systems*, Seattle, WA: Association for Computing Machinery, 2001, pp. 490–497.
[12] M. A. Chakravarthy, "Mining the semantic web," in *Proc. Workshop at the 13th European Conference on Machine Learning/6th European Conference on Principles and Practice of Knowledge Discovery in Databases*, vol. 4, no. 2, 2005.
[13] J. Davies, D. Fensel, and F. V. Harmelen, *Towards the Semantic Web*, New York, NY: John Wiley and Sons, 2002.
[14] T. B. lee, J. Hendler and O. Lassila, "The semantic web," *Scientific American*, 2001, vol. 284, no. 95, pp. 34–43.
[15] K. Abdulnazeer and P. Abdulhaleem, "Standardization of unstructured textual data," in *Proc. the 2nd Indian International Conference on Artificial Intelligence*, 2005, Pune, India, 20–22 December 2005.
[16] S. Kampa, T. M. Board, L. Carr, and W. Hall. (2001). Linking with meaning: Ontological hypertext for scholars. Technical report. Southampton: University of Southampton. [Online]. Available: http://www.bib.ecs.soton.ac.uk/data/5163/pdf/lwm.pdf
[17] O. Lassila, "Web metadata: A matter of semantics," *IEEE Internet Computing*, vol. 2, no. 4, pp. 30–37, 1998.
[18] C. C. Aggarwal, "Collaborative crawling: Mining user experiences for topical resource discovery," IBM Research Report, 2002.
[19] C. Kemp and K. Ramamohanarao, "Long-term learning for web search engines," in *Proc. European Conference on Principles and Practice of Knowledge Discovery in Databases*, 2002, pp. 263–274.
[20] S. Downes, "Learning objects: Resources for distance education worldwide," *International Review of Research in Open and Distance Learning*, vol. 2, no. 1, 2001.
[21] J. Quemanda and B. Simon, "A use-case based model for learning resources in educational mediators," *Educational Technology and Society*, 2003, vol. 6, pp. 149–163.
[22] M. Rouse, *Essential Guide: Text Mining (Text Analytics)*, Search CRM, 2010.
[23] L. Stojanovic. (2013). eLearning based on the semantic web. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.16.295&rep=rep1&type=pdf
[24] D. Maynard, *Text Mining and the Semantic Web*, Manchester: University of Manchester, 2005.

Hamad Ibrahim Alomran was born on 12 September, 1969, Riyadh, Saudi Arabia, He is currently an associate professor of Information Management at College of Computer and Information Sciences, Al-Imam Muhammad ibn Saud Islamic University, where he teaches and does research.

Alomran received his master degree from North Carolina Central University of USA, and his Ph.D. from Imam University of KSA.

His research interests have focused on information security awareness, smartphones information retrieval, information behaviors, information architecture, data mining.

He is now working as a consultant to the rector of Majmaah University, and the dean of Prince Salman Institute for Studies and Consulting Services.