Information and Knowledge Management
ISSN 2224-5758 (Paper) ISSN 2224-896X (Online)
Vol.7, No.9, 2017

Development of an Enhanced Knowledge Retrieval System Using Web 2.0 Technology and Vector Space Model

Dr. Bola Oladejo[1*] David Awolola[2]
1.Senior Lecturer, Computer Science Department, University of Ibadan
2.Masters Student, Computer Science Department, University of Ibadan

Abstract
There is an increasing pool of information on the web and a major contributor is web 2.0 technology on which social media is based. Searching for specific information in this pool is always tasking, therefore, the need to harness this information as a means of enhancing retrieval and reuse of relevant ones. Some researches and development have been carried out in the field of Knowledge Retrieval using Vector Space Model (VSM) and Latent Semantic Indexing (LSI), but the approach used is based on large pool of information available online, which makes getting most relevant information relatively difficult at the point of retrieval, this is a major setback. Collaborations on Facebook and Twitter (web 2.0 technology) were harvested using APIs and stored in the Knowledge Repository, The collaboration on social media served as the source of information in the Knowledge Repository. An Enhanced Knowledge Retrieval System (EKRS) applying VSM was developed and implemented. The use of VSM was to calculate the Cosine Similarity and Term Frequency to aid effective retrieval of relevant documents from the repository based on user's needs. In this project, we were able to achieve the aim of retrieving relevant documents. EKRS was able to employ both web 2.0 and VSM to meet specific user's information needs.
Keywords: web 2.0, Knowledge retrieval, Vector Space Model, Latent Semantic Indexing, Knowledge Repository, Cosine Similarity and Term Frequency.

1. Introduction

On a daily basis, information on internet is increasing. The internet is a central source of information for researchers and discussion groups. For researchers and other individuals that visits the internet every day to source for information, getting the right information has been a challenge because of the abundance of both relevant and irrelevant information on the internet in congruence to the users' information need. A major contributor to the abundant information on internet is web 2.0 technology. Web 2.0 technology is a technology on which social media sites are based, e.g. Facebook, Twitter, blog, LinkedIn etc. It is a new generation of Internet-based collaborative tool, that has increased in popularity, availability, and power in the last few years (Kane & Fichman, 2009). Web 2.0 is a set of Internet-based applications that harness network effects by facilitating collaborative and participative computing (O'Reilly, 2006). This website gives room for interaction and collaboration, idea sharing, socialization with friends and contribution to other people's post among others. As individuals and groups share idea on this platform on a daily basis, it increases the abundance of information on the internet, thereby giving access to more information when needed. Processed information is what results to knowledge and Knowledge is becoming strategically important resources and a very significant driver of organizational performance (Yesil & Dereli, 2013). For this knowledge to be useful for any individual or organization growth then there is need for proper management of such knowledge else they become useless. This then leads to Knowledge Management. "KM describes the processes of acquiring, developing, sharing, exploiting and protecting organizational knowledge to improve organization's competitiveness." (Gaal et al., 2015).

The activities involved in KM are: knowledge capture and/or creation, knowledge sharing and dissemination and knowledge acquisition and application (Dalkir, 2005). Knowledge acquisition and application can also be referred to as knowledge retrieval and reuse. Knowledge retrieval can be defined as searching for materials (usually documents) of structured or unstructured nature (usually text) which satisfies an information need from with a large collection from different source (usually store in the database) and return the knowledge in a structured form. (Christopher, Prabhakar & Hinrich, 2009). Knowledge retrieval is built on information retrieval process. The difference is that the knowledge retrieval deals with structured and classified information which has been put in different domain. The goal of knowledge retrieval is to reduce the burden of going through and analyzing a long list of data set or documents unlike Data Retrieval System (DRS) or Information Retrieval System (IRS).

Web 2.0 technology avail users the opportunity to post information, collaborate with others that share common interest and reach out to distant people. But this technology only connects people and not information. The information shared on this platform either by individuals or a community of people with common interest are not well organized this makes accessibility to such information in relevant manner to specific users' need is limited.

Other retrieval system have a large pool of information in their repository, hence it cannot optimally satisfy the specific users need of relevant document to their information need. Hence the need for a system that deals with specific information and area of interest that will satisfy users need.

This paper developed and implemented a system that effectively help users retrieve relevant document to their information needs.

The rest of this paper is divided into four sections. The theoretical background and review of related works is in section 2. Section 3 reports the research methodology while section 4 discuss the system implementation and results. The paper is concluded on section 5.

2. Theoretical Background

2.1Knowledge

Knowledge must be distinguished from information and data so as to get a true picture of it in Knowledge Management (KM). Data is raw facts or figure that make little or no meaning to users. While information is processed data for proper usage. According to Davenport and Prusak (1998) "knowledge is a fluid mix of framed experience, values, contextual information, expert insight and grounded intuition that provides an environment and framework for evaluating and incorporating new experiences and information". Knowledge is sensitive making it difficult to capture in words or have a complete understanding of it logically. Knowledge is part and parcel of human makeup therefore it is hard to pin knowledge down in a place or express and articulate it. Knowledge in higher institution of learning is neither individually owned nor static, but embedded in individual employees or student of the institution (academic staff, non-academic staff, students and top management), Project teams, faculty and university. But knowledge is very important in decision making for any organization to compete with her counterpart. Knowledge in the context of this work can be refer to as contributions of individual in a group discussion driven towards solving a particular problem. It is represented in form of written text that others can see, Knowledge can be generally categorize into two which are:

Tacit Knowledge: it is any knowledge that has not been explicitly represented or articulated in any form; therefore, it cannot be stored, retrieved, copied and transferred because it is still in the individuals that possesses them. (Fleck, 1996). This knowledge is part of an individual that possess them, it cannot not be seen or understand by others except it is expressed by such individual.

Explicit Knowledge: this type of knowledge can be expressed in words or symbol like image representation; therefore, it can be stored, retrieved, copied and transferred into a written document which can be used at any other time when the need arises. (Hansen 1999). This knowledge no longer belong to an individual, others now have access to it and can make use of it in for other problem solving activities.

2.2Knowledge Management

The origins of Knowledge Management (KM) can be traced back to the late 1970s. Knowledge is very significant and important for high performance of any institution today (Yesil & Dereli, 2013). KM is a very important tool for any organization, institution, industry and government today for competitive advantage, innovation and expansion. In defining KM with respect to this work, it will be necessary to incorporate how the knowledge is captured, stored and of what value it is. In knowledge management efforts have been placed on various aspects which include knowledge capturing, codifying and sharing the knowledge with other people. To define KM therefore according to Rigby (2009) as "Knowledge management develops systems and processes to acquire and share intellectual assets. It increases the generation of useful, actionable, and meaningful information, and seeks to increase both individual and team learning. In addition, it can maximize the value of an organization's intellectual base across diverse functions and disparate locations". KM is on the increase today so much that companies are now using it as a leverage for competition based on what they know and how they are able to efficiently use that knowledge and acquire new ones (Davenport & Prusak, 1998). KM afford any institution to keep abreast latest discoveries and to move with the trend so as not to remain obsolete in a dynamic world. Creation, transfer or sharing and use of knowledge has become an increasingly important factor for organization competitiveness. KM gives room for knowledge creation then management. An environment which encourages knowledge to be created, shared, enhanced, organized and utilized for the benefit of the users.

2.3Knowledge Retrieval (KR)

Knowledge retrieval is one of the major components of KM System. The goal of KR is to satisfy the users with documents relevant to their query term on the web. To achieve this goal and to ascertain the retrieval efficiency, documents are transformed into suitable representations that are retrievable. KR is a combination of Data Retrieval System (DRS) and Information Retrieval System (IRS). DRS uses relevant data to acquire knowledge, the problem to be solved are well structured and concept definitions are clear while IRS uses relevant information to acquire knowledge, the problem is semi-structured and concept definitions are not always clear (Yao, Zeng, Zhong&Huang, 2007). Knowledge Retrieval focuses on how to extract, represent and use the

knowledge in data and information. This retrieval system provides knowledge to users in a structured form (Bellinger, Castro & Mills. 2004).

2.3.1Knowledge Retrieval Model

In this age, information on the internet is enormous, widely-distributed, semi-structured and interconnected that it is sometimes difficult to find a relevant information to that one needed by the user. A major goal of knowledge retrieval process is to provide different users with relevant document, information or knowledge that will satisfy their need. There are different models used for information retrieval among which are1. Set-theoretic models which is sub-divided into Boolean Model and Fuzzy retrieval; 2. Algebraic models subdivided into Vector Space Model and Latent Semantic Indexing. 3. Probabilistic Models.

Boolean Model: Boolean model is the simplest of other retrieval methods that relies on the use of Boolean operators (AND, OR and NOT) and classical set theory that exists in the documents to be searched and the user's query are conceived as set of terms. It is easy to implement but construction of effective Boolean queries is difficult. During knowledge retrieval phase, queries are less than perfect in two respects which are: first, they retrieve some irrelevant documents. Second, they do not retrieve all relevant documents.

Latent Semantic Index (LSI): Matching of terms in documents with that of the query is a process of information retrieval but lexical matching methods can be inaccurate when they are used to match a user's query because a word can be expressed with much different vocabulary (synonyms). These may affect the effectiveness of query search given by the user because many documents that do not bear the literal term in a user's query may not be match. Latent Semantic Indexing (LSI) is an information retrieval model that tries to overcome the problems of lexical matching. In order to achieve this, it uses statistically conceptual indices instead of individual words comparison for retrieval used by other technique.

Vector Space Model (VSM): VSM is used for information filtering, retrieval, indexing and ranking of relevant documents. VSM process can be divided into three stages which are: (1) Document indexing; here all document that bears the query term are extracted from the pool of other document text. (2) In this stage, the weighting of the indexed terms in step one above is carried out so that the retrieval process can be enhanced and relevant documents are retrieved. (3) After the retrieval of relevant documents is done, they are ranked with respect to the query according to a similarity measure.

2.4Knowledge Management Using Web 2.0

The main aim of web 2.0 is to improve collaboration and interaction. This interaction and collaboration is also an important aspect of Knowledge Management (KM) called "Socialization" approach. Therefore, it can be deduced that when web 2.0 application is used for KM, it will both support and improve knowledge sharing and creation. Application of web 2.0 to KM can potentially lead to a new era of KM known as KM 2.0. (Bebensee, 2010). KM tools are technology that enhance and enable knowledge acquisition, codification, transfer and realization (Egbu & Botterill, 2002). A major and critical aspect of web 2.0 is that it ensures rich user experience and it also contribute immensely to information exchange through it rich peer-to-peer user interactions, it support collaborative value creation and combine the best elements of KM, which includes suitability for business environments and overcomes lots of setbacks like limited opportunities for simultaneous collaboration (Wagner & Majchrzak, 2006 in Nath, 2012).

Therefore, web 2.0 has a great potential to solve one of the great challenges of KM by capturing tacit knowledge and converting it into explicit knowledge. It is also discovered that web 2.0 have the ability to combine traditional KM tools features with social computing where knowledge is evolved through social interaction (Parameswaran, 2007).

2.5Review of Related Works

Chi-un lei et al., (2012) explored the possibilities of using social media networking tools to support teaching practice in technological courses. Facebook page (for content sharing) and blog (as a tool to express thoughts and opinion) were used as web 2.0 tools for this research. They discovered that Facebook is more effective for teacher-to-student communication because facebook page can be used as a complementary tool to announce course news, share course materials and stimulate students' informal discussions while blog is effective for collaborative writing.

Boubekeur & Azzoug (2013) discovered that Information Retrieval System (IRS) rely on keywords to index documents and queries, documents are retrieved based on the number of shared keywords with the query. Semantic-focused retrieval approach to overcome the keyword problem was proposed. The authors aim to retrieve documents that are semantically relevant to a given user query. To achieve this, they carried out an experiment using a dataset that contains 423 documents consisting of newspaper articles from the TIME magazines. 83 queries were performed. The work enabled retrieval of documents that are semantically relevant to a given user query. This work did not put into consideration the similarity measure score which aids most relevant retrieval in Vector Space Model (VSM)

Kulaki & Mahony (2014) viewed web 2.0 as an innovative communication platform that has encouraged people to share their thoughts and experiences in a collaborative way. Top 50 UK institutions and top 50 world institutions of learning were compared to see how many of the institutions are on social media. The survey studied how many of these institutions use it as a means of collaboration and which of the web 2.0 tools is mostly accepted and used. It was discovered that web 2.0 has brought new dimension to teaching and learning and that Facebook, twitter, YouTube are common tools used in most of these institutions.

Manwar et al., (2012) used Vector Space Model (VSM) for information retrieval because of its advantage over the Boolean technique. Boolean technique is a lightweight model that matches query with precise semantics and this can cause the results to be tides and thereby missing partial matching. The goal of this research was to retrieve the most relevant documents. VSM was used with Matlab on Cranfield data collection of aerodynamics domain. It was discovered that it was able to recover about 89.41% relevant documents.

Bebensee (2011) carried out a research on the usage of web 2.0 technology in organizations so that collaboration can be boosted thereby increasing the level of innovation among workers. Based on the research carried out by the author, web 2.0 can be thought of as some virtual communities that facilitate the sharing of information and knowledge with the web. The author indicated that there is need for organizations to recognize knowledge as a resources. But the author only focuses on organizations and the need for them to adopt web 2.0, did not put into consideration the university community and which of the web 2.0 application will serve them best.

3. Enhanced Knowledge Retrieval System (EKRS)

Figure 3.1 presents a comprehensive outline of different stages involved in the development of Enhanced Knowledge Retrieval System (EKRS). This framework highlights the process and components for analysis, design and development of the system. It also shows the life cycle of EKRS which is in three phases:

1. Knowledge Creation: This is the point of collaboration and sharing on web 2.0 technology i.e. social media platforms.
2. Knowledge Storage: Harvested knowledge from social media platforms are stored in the knowledge repository.
3. Knowledge Retrieval: Vector Space Model is applied on stored knowledge for effective knowledge retrieval and reuse.

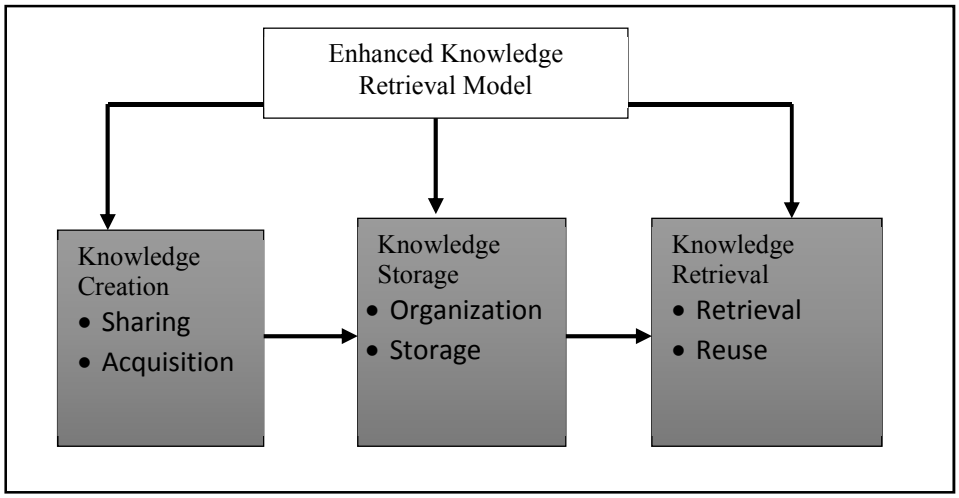

Figure 3.1: Framework of Enhanced Knowledge Retrieval System

Vector Space Model represents the documents and queries as vectors in a multidimensional space, whose dimensions are the terms used to build an index to represent the documents. The document of interest in this work was captured from the discussion forum on different web 2.0 platform which was organized and stored in a knowledge repository. The document vector is the size of all document present in the repository, while the query vector is the size of the Query Term (QT) as written in the query box by the user. The similarity function between document vector Di and query Q is represented as:

$$\text{Cosine}\theta = \text{Sim}(Q, D_i) = \frac{\sum_{i=1}^{V} W_{Q,i} \times W_{d,i}}{\sqrt{\sum_{i=1}^{V} W_{Q,i}^2} \times \sqrt{\sum_{i=1}^{V} W_{d,i}^2}} \quad \text{equation 1}$$

Where Q represents Query Term
i represents term either in document or in query
D_i represents number of documents or domain

$W_{Q,i}$ represents the weight term i in the query
$W_{d,i}$ represents the weight term i in the document

$$W_{di} = tf_{di} \times idf = tf_{di} \times \log\frac{D}{df_i} \quad \text{equation 2}$$

IDF = $\log(\frac{D}{df_i})$ where df_i represents number of documents containing term i.

4. Implementation and Results

Enhanced Knowledge Retrieval System is developed and tested to ascertain its functionality and measure its effectiveness.

4.1 Result Generated using Vector Space Model for Retrieval

Vector Space Model view the distance between documents as vector space and calculate the cosine similarity score and the term frequency to determine relevancy of such documents to the query. We have 28 documents extracted from different pages on twitter and Facebook and they are stored in the repository. Performing a knowledge retrieval on the documents in the repository. We perform 10 different search with different words and phrase using VSM, we were able to calculate the cosine similarity for each query to the documents. Table 4.1 shows the query represented by Qi to Qn and the Cosine Similarity Score.

Table 4.1: Cosine Similarity Score of Query and Document

Query	Max Cosine Similarity Score	Min Cosine Similarity Score	Avg Cosine Similarity Score
Q_1	0.52	0.16	0.34
Q_2	0.19	0.002	0.10
Q_3	1.0	0.24	0.62
Q_4	0.19	0.006	0.10
Q_5	0.15	0.46	0.31
Q_6	1.0	0.86	0.93
Q_7	1.0	0.58	0.79
Q_8	0.22	0.04	0.13
Q_9	0.22	0.05	0.14
Q_{10}	0.09	0.02	0.06

4.2 Evaluation of Relevant Documents Retrieved using Precision, Recall and F-measure

In information retrieval, relevance of retrieved documents or its irrelevance is a very important thing and to calculate this we use precision, recall and F-measure. Precision, recall, and the F-measure are set-based measures. They are computed using unordered sets of documents in a repository or file. Precision is the fraction of retrieved instances that are relevant, while recall is the fractions of relevant instances that are retrieved. F-measure combines precision and recall by calculating the harmonic mean of precision and recall. After 10 search were perform, precision, recall and F-measure were calculated and table 4.2 shows the result of the calculation.

Table 4.2: Precision, Recall and F-Measure value

Query	Precision	Recall	F-Measure
Q_1	0.6	1.0	0.75
Q_2	0.3	0.6	0.40
Q_3	0.2	0.3	0.24
Q_4	0.3	1.0	0.46
Q_5	0.3	0.4	0.34
Q_6	0.2	0.3	0.24
Q_7	0.5	0.4	0.44
Q_8	0.8	0.7	0.75
Q_9	0.5	0.4	0.44
Q_{10}	0.8	0.6	0.69

5. Conclusion

The ability to harvest knowledge from social media platforms to populate the database was achieved using the API for each social media platform. This knowledge was then classified into various group of relevance and stored in the knowledge base. Although different information retrieval system has existed before now but this system was narrowed down to a special purpose Knowledge Retrieval System, with retrieval based on collaborations among those that shares common interest in their chosen field. Using the precision and recall for evaluation including cosine similarity measure, it became clear that the retrieval system is efficient and effective

having 92.59% relevant documents retrieved during the testing phase.

References

Babensee, T, Helms, R and Spruit, M. (2010), "Exploring Web 2.0 Applications as a Mean of Bolstering up Knowledge Management" *The Electronic Journal of Knowledge Management Volume 9 Issue 1 (pp19)*, available online at www.ejkm.com, Accessed November, 2016)

Bellinger, G., Castro, D. and Mills, A. (2004) Data, Information, Knowledge, and Wisdom, http://www.systemsthinking.org/dikw/dikw.htm (accessed January 24th, 2017).

Boubekeur F. and Azzoug W. (2013), Concept-Based Indexing in Text Information Retrieval. *International Journal of Computer Science & Information Technology (IJCSIT) Vol 5, No 1*

Chi-Un, L., et al. (2012), Using Web 2.0 Tools to Enhance Learning in Higher Education: A Case Study in Technological Education. *Proceedings of the International Multi Conference of Engineer and Computer Scientists 2012 Vol II, Hong Kong*.

Christopher D. M., Prabhakar R., Hinrich S., (2009) An Introduction to Information Retrieval. *https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf* pp 38 (Accessed October, 2016)

Dalkir K. (2005), Knowledge Management in Theory and Practice. *Elsevier Butterworth–Heinemann, Copyright © 2005, Elsevier Inc. All rights reserved. pp 26-44*

Davenport, T. H., & Prusak, L. (1998). Working Knowledge: How Organizations manage what they know. *http://www.kushima.org/is/wp-content/uploads/2013/09/Davenport_know.pdf* (Accessed May, 2016)

Egbu C. O. & Botterill K. (2002). Information Technologies For Knowledge Management: Their Usage And Effectiveness. ITcon, Vol. 7; *http://www.itcon.org/2002/8* pp. 125-137

Fleck, J. (1996). Informal information flow and the nature of expertise in financial services. *International Journal of Technology Management, 11(1-2), 104-128*

Gaál, Z., Szabó, L., Obermayer-Kovács, N. and Csepregi, A., (2015), "Exploring the role of social media in knowledge sharing". *The Electronic Journal of Knowledge Management Volume 13 Issue 3 (pp185-197) available online at www.ejkm.com*

Hansen, M. T., Nohria, N., & Tierney, T. (1999). What's Your Strategy for Managing Knowledge? *Harvard Business Review, 77(2), 106-116. https://hbr.org/1999/03/whats-your-strategy-for-managing-knowledge*

Kane, G. C., & Fichman, R. G. (2009). The Shoemaker's Children: Using Wikis for Information Systems Teaching, *Research, and Publication. MIS Quarterly, 33(1), 1–17. http://dl.acm.org/citation.cfm?id=2017412*

Kulakli, A. and Mahony, S. (2014), Knowledge creation and sharing withWeb 2.0 tools for teaching and learning roles in so-called University 2.0. *Elsevier Procedia - Social and Behavioral Sciences 150 (2014) 648 – 657. 10th International Strategic Management Conference.*

Manwar, A. B., Mahalle, H. S., Chinchkhede, K. D. and Chavan V. (2012), A Vector Space Model for Information Retrieval: A Matlab Approach, *Indian Journal of Computer Science and Engineering (IJCSE) Vol. 3 No. 2.*

Nath, A. K., (2012), "Web 2.0 Technologies for Effective Knowledge Management in Organizations: A Qualitative Analysis", *https://libres.uncg.edu/ir/uncg/f/Nath_uncg_0154D_10898.pdf* (Accessed April 2016).

O'Reilly, T. (2006). Web 2.0 compact definition: Trying again. *http://radar.oreilly.com/archives/2006/12/web-20-compact.html (Accessed October 2016)*

Parameswaran, M. (2007) "Social Computing: An Overview", Communications of AIS, (19), pp.762-780. *http://people.sunyit.edu/~krieseg/Scrapbook/data/20111105161503/105_social_computing_an_overview_234.pdf (Accessed September, 2016)*

Rigby, D. 2009 Management Tools 2009: An Executive's Guide, *http://www.bain.com/management_tools/home.asp.* (Accessed May, 2016)

Yao Y., Zeng Y., Zhong N., and Huang X. (2007), Knowledge Retrieval, *IEEE Xplore Digital Library (Accessed January, 2017)*

Yesil, S. and Dereli, S. F. (2013), An empirical investigation of organizational justice, knowledge sharing and innovation capability. *SciVerse Science Direct, vol. 75, pp. 199-208*.