Supporting Personalised Content Management in Smart Health Information Portals

Daswin De Silva
Frada Burstein
Julie Fisher

Centre for Organisational and Social Informatics
Faculty of IT
Monash University
Victoria, Australia
Email: daswin.desilva@monash.edu

Abstract

Information portals are seen as an appropriate platform for personalised healthcare and wellbeing information provision. Efficient content management is a core capability of a successful smart health information portal (SHIP) and domain expertise is a vital input to content management when it comes to matching user profiles with the appropriate resources. The rate of generation of new health-related content far exceeds the numbers that can be manually examined by domain experts for relevance to a specific topic and audience. In this paper we investigate automated content discovery as a plausible solution to this shortcoming that capitalises on the existing database of expert-endorsed content as an implicit store of knowledge to guide such a solution. We propose a novel content discovery technique based on a text analytics approach that utilises an existing content repository to acquire new and relevant content. We also highlight the contribution of this technique towards realisation of smart content management for SHIPs.

Keywords

Smart health information portal, personalised content management, automated content discovery, text mining, vector space model, query extraction, content discovery and ranking.

INTRODUCTION

An information portal, in general, is a gateway to a diverse collection of information on a specific domain of interest. It attempts to aggregate information from multiple sources and present these in a useful form to targeted groups of users (Collins 2002, Tatnall 2005). A health information portal (HIP) follows a similar model of operation with a high degree of emphasis on relevance, usefulness, reliability and timeliness (collectively identified as quality) of information due to its crucial role in human wellbeing (Xie and Burstein 2011). Personalisation of content is equally crucial to a HIP as its audience will be composed of several sub-groups with diverse interests, needs and expectation (Fisher et al 2007). The credibility of a HIP as a useful information resource is directly associated with the quality of its content. Therefore, to ensure the quality of information delivery, it is imperative for domain experts to examine all content delivered by a HIP.

Portal technology has made inroads into all key sectors of an economy including health, education, government and commerce. Predictions for e-health (Mandl et al 1998) and use of portal technology (Collins 2002) from a decade ago have been realised in recent times with the widespread implementation and adoption of portals to promote health and wellbeing (Theofanos and Mulligan 2004, Kukafka et al 2007). Similar outcomes are evident in other sectors; e.g. government (Elmagarmid and McIver 2001) and education (Katz 2002). The advent of portable smart devices and expectations of the Semantic Web further enhance the value and need for continuing development in portal technologies.

Advances in information systems coupled with the wide availability of diverse interfaces to the Internet have led to the adoption of smart technology for the development of portals. Within this context, it is pertinent to formally define a smart health information portal (SHIP) as the provision of smart technology and techniques to enhance the core capabilities of content management, content delivery and collaboration. We argue that it is not sufficient to define SHIP based exclusively on its exhibiting computational intelligence features, e.g. learning, reasoning and memory. Sustainability of SHIP operation within organisational settings is crucial for its long-term viability. Hence, the issue of maintenance support becomes one of the deciding factors in the level of *smartness* of a SHIP's operation.

Breast Cancer Knowledge Online (BCKOnline, www.bckonline.monash.edu.au) and Heart Health Online (http://www.sip.infotech.monash.edu.au/heart-portal/) are examples of SHIPs researched and developed at the Faculty of IT, Monash University to address the health and medical information requirements of individuals associated with breast cancer and mental health associated with heart conditions, including patients, carers, family and friends of those affected. The delivery of user-sensitive, relevant, timely and accurate health information to the various user groups was the focus throughout the various phases of the projects. These SHIPs are implementing several novel research outcomes, e.g. resource description quality criteria modelling (McKemmish et al 2009), user-centric portal design (Fisher et al 2004), automated quality assessment (Xie and Burstein 2011) and decision support systems perspective on portals (Burstein et al 2005). Reported experience from the development of these SHIPs clearly demonstrated the value of continuous engagement and a high degree of reliance of user groups to identify, categorise and describe the type of information required by relevant individuals. The resource intensity in terms of time and scarcity of relevant expertise was also highlighted by the researchers involved in these projects (Burstein et al 2005, Burstein et al 2006, Pier et al 2008). These studies reinforce the need for intelligent support for SHIP content management.

Automated content discovery (ACD), content summarisation (Erkan and Radev 2004), dynamic ranking, user annotations and feedback (Ciccarese et al 2011) are some of the enhancements to content management, which could assist in SHIP content management. Content delivery is enhanced with user profiling, geographical filtering, mobile interfaces and device independent content delivery. Online messaging, social networking and discussion forums are enablers for smart collaboration. Among these features, assurance of quality of information delivery is by far the most sought after by users, and the most resource intensive from the organisational set up point of view. Based on the aforementioned definition of resource quality, SHIP content management can be effectively supported with the help of smart technology. Text analytics is an emerging area in business analytics where smart techniques are being developed and used to extract patterns, predictions and semantic content from text corpora (Aggarwal and Zhai 2012). This paper proposes a novel ACD technique for a SHIP (ACD-SHIP) based on a text analytics approach to assist domain experts to acquire new resources for the content repository.

In order to apply this technique it is necessary to understand the content management lifecycle and its composition. Section Two presents these two aspects of content management followed by a background study on related areas of research and outcomes. Section Three explicates the ACD-SHIP technique, its conceptual basis and functionality. The proposed ACD technique was applied to the BCKOnline knowledge repository data with positive results indicative of its role in supporting personalised content management. Section Four exhibits the outcomes from this application. A discussion on improvements to the ACD technique and overall developments in smart technology for HIPs content management capability sees to the end of the paper.

CONTENT MANAGEMENT IN A SHIP

Content management (CM) is a widely published topic with research conducted in knowledge management (Bonifacio et al 2008), Internet research (Tatnall 2005) and information retrieval (Bates 2011). The focus of research in content management is largely influenced by its context. This context varies from enterprise level management to management of basic website content. At the enterprise level, recent advances include the ECM3 model (ECM3 2009) which aims to address the CM challenges by introducing stages of maturity for all enterprise documents and unstructured content. The web content maturity model proposed by Forrester research (WCCM 2009) attempts to address the challenges facing an organisation’s web content. It consists of four phases; basic, tactical, enterprise and engagement. The focus gradually broadens through these four phases, starting with the basic focus of making enterprise content available online and in the final phase expanding it to providing an online channel to achieve organisational goals. Boiko, (2005) defines CM as composed of three phases, the first is creation or collection of content, the second phase is managing storage and retrieval, versioning over time and multiple languages etc. The third phase involves publication and delivery of the content.

However, content management in SHIP (and information portals in general) is distinguished from content delivery as these two features along with collaboration form the three core capabilities of a portal (Chau et al 2006). This distinction is justified by the difference of scope in content management of an organisation/e-business with that of an information portal, particularly because the latter is conceptually centred on content. As user-sensitive content delivery is identified as a core feature of SHIP, content management becomes one of the key determinants of its success and long term viability.

A SHIP does not create content, it acquires content from other creators and introduces a layer of personalisation atop the content. This layer of personalisation encapsulates the domain expert’s knowledge and awareness of both the health issue(s) and the target audience (Burstein et al 2005). This is another reason to distinguish between the significant roles played by content management, delivery and collaboration for the successful adoption of a SHIP by its target audience. The following sections elaborate on the CM lifecycle and the CM model for SHIPs.

Content Management Lifecycle

The process of content management for SHIPs is cyclic primarily due to the dynamic nature of health information and the requirement for maintaining quality of advice. It is crucial for a SHIP to maintain relevant and timely health information; thereby regular revision of all content is a stringent indicator of quality of a SHIP. As illustrated in Figure 1, the main phases of CM are locate, personalise, store and revise. The diagram also depicts the contributing elements that flow from one phase to the other.

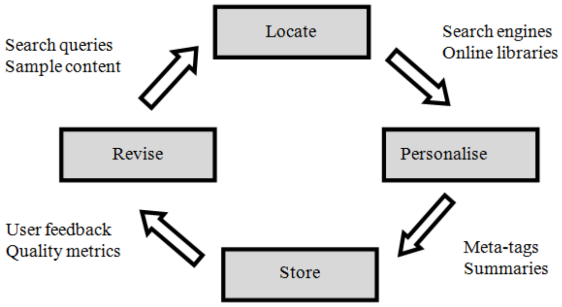


Figure 1: Content management lifecycle

Locate – This is the initiating phase of the lifecycle. The domain expert identifies relevant terms, phrases and any other resource identifiers which are used to query search engines, online libraries, academic journals and other portals to search and acquire appropriate content.

Personalise – Personalisation is the process of refining the acquired content. As mentioned earlier, this layer captures and applies a domain expert's knowledge to the content. Meta-tags are introduced to identify which audiences would find the information relevant and useful. Content summaries are compiled to make search results manageable and certain resources meaningful. Quality metrics are introduced to convey credibility of the content creator to the end-user.

Store – Storage deals with maintenance of the acquired content and personalisation attributes in a structured manner that can be useful for fast delivery and convenient revision. It is also necessary to store user feedback (and optionally comments) for each resource. Feedback is a direct indicator of quality as perceived by the target audience.

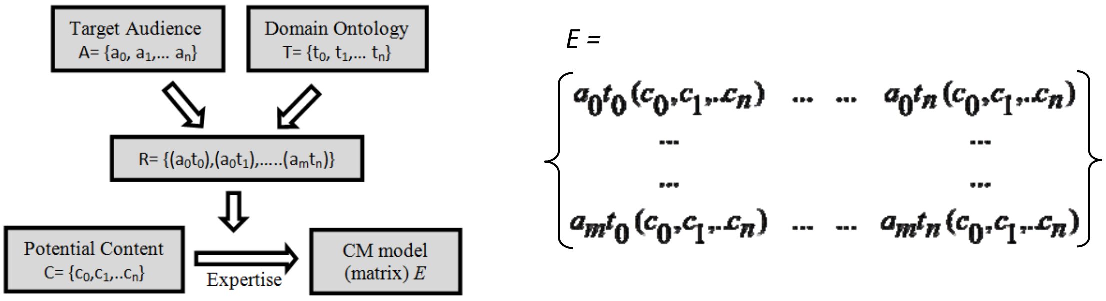
Revise – Revision of content is crucial to a SHIP as health information is prone to frequent changes due to the fast-paced nature of research in health. Revisions take into account user feedback as well as quality metrics. It measures relevance and timeliness of a resource after a lapse of time. The *Revise* phase completes the cycle by providing search queries and sample content (of which more is required) to the locate phase.

The concept of content personalisation is the major expected feature, which differentiates portals from generic websites (Collins 2002). Personalisation is practiced in several areas of research including user profiling, recommender systems and information retrieval. In Nam-Kim et al (2011) a hybrid recommender approach that uses both collaborative and content filtering to improve the level of content personalisation is proposed. User profiling attempts to generate models of user sub-groups, commonly based on user behaviour. Middleton et al (2004) propose a recommender system for a research database based on ontological user profiling. In Abbar et al (2004), the authors present a personalized access model that provides a generic set of concepts and techniques, which can be deployed over a given architecture to make applications adaptable to users' profiles and contexts. It is acknowledged that a fair amount of domain expertise and/or user feedback is needed in order to implement a good level of personalisation in a portal. Overall, existing approaches to content personalisation are at the two ends of collaborative filtering, content filtering, or a combination of both. The ACD technique suggested in this paper follows a content filtering approach when acquiring new personalised content for the repository. We also note that the content discovered needs to go through a process of validation before it can be incorporated into the SHIP. Such validation requires involvement of a domain expert or a process guided by well documented domain expertise. In both cases, it results in semi-automated process of content management, which supports a more efficient operation of a SHIP.

SHIP Content Management Model

The CM model represents the external entities of content management and their interactions in the formulation and management of personalised content. Informed by the experience with BCKOnline and Heart Health portal research (Pier et al 2008), this model is a conceptualisation of the fact that the audience of the SHIP users has distinct characteristics and contexts, which potentially affect their information needs. The resources for a SHIP can be aligned with a domain ontology, which classifies them against the major concepts, which define such a domain. For example, official publications from medical journals are usually classified by a set of key words, which the audience is likely to use to search and retrieve these publications. A set of such key words or subject terms can be considered as part of domain ontology. The completeness or relevance of such an ontology can be problematic especially when it comes to the search for relevant user-centred information (Burstein et al, 2005), however, these issues are outside the scope of this particular paper. For this research we assume that there is a trusted and appropriate domain ontology constructed for resource classifications (for example, in BCKOnline a combination of Medical Subject Headings (MeSH), BreastCare Victoria Glossary, BCKOnline Disease Trajectory and BCKOnline 'Key Words' were used as encoding schemas for the Subject metadata element (see: http://infotech.monash.edu/research/about/centres/cosi/projects/bcko/about.html). The role of the Domain Expert in classifying potential resources against the needs of the target audience becomes essential for identifying the best terminology suitable and understandable by the target audience.

At the generic level, the target audience, potential content, a domain ontology and domain expertise are the external entities that are fused together to generate personalised content. This formulation is further illustrated in Figure 2a. It is useful to formally define the entities and their interactions. The target audience comprises sub-groups of users with similar characteristics and thus having similar information needs from the SHIP. Let $A= \{a_0, a_1,\ldots a_n\}$ be the target audience comprising all sub-groups. Health resources are all content from a particular health domain that can be made accessible through the SHIP to the target audience. Let $C= \{c_0, c_1,\ldots c_n\}$ be the set of all such content. A domain ontology formalises the concept hierarchy of knowledge for a specific domain, it can be represented, simply, as a set of topics, $T= \{t_0, t_1,\ldots t_n\}$. The information requirements for audience A is determined using the Cartesian product of A and T. Let R be the Cartesian product, $R=A*T$. Actual information requirements could very well be a subset of R as all terms may not be applicable to all A. Domain expertise transforms information requirements R, to actual content C, by determining subsets of C that address each R. Let this transformation be $E = \{e_0, e_1,\ldots e_n\}$, where $e_0 = \{a_0t_0,(c_0,c_1,\ldots c_n)\}$ comprises information requirements and a set of matched content elements. The transformation E represents the CM model as it captures all entities and their relationships. It can also be visualised as a matrix (Figure 2b).



(a) (b)

Figure 2: (a) Formulation of content management entities (b) SHIP content model as a matrix

The CM model possesses certain properties that make it robust and flexible to changes. Over time, it is likely A, T and C would expand or contract to reflect developments in health practices. Matrix E is time-invariant and thus can be altered easily to reflect these changes. The challenge and opportunity for developing a sustainable SHIP is in designing transformation R as a semi-automated expert-driven procedure by using intelligent technologies. In the next section we describe how such procedure can be enhanced by using automated content discovery.

AUTOMATED CONTENT DISCOVERY FOR SHIP (ACD-SHIP)

The proposed technique, termed the ACD-SHIP technique is delineated in this section. It enhances the content management capability of a SHIP with the use of smart technology for the content revision phase. The CM model is the main source of information for the proposed technique. It extracts semantics that are useful to formulate queries that discover new content as well as semantics that are used to measure the relevance of new content from the CM model. The new content is reviewed by an expert prior to inclusion in the repository. The workflow of the proposed technique is illustrated in Figure 3.

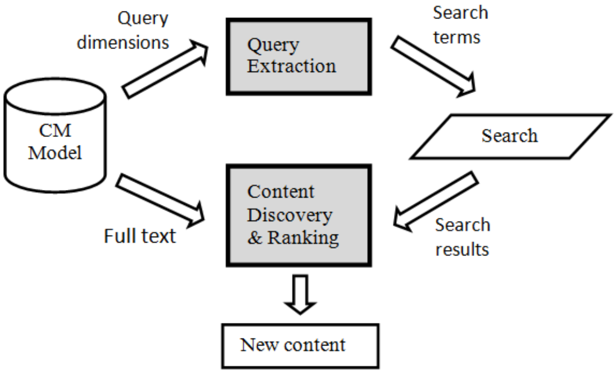


Figure 3: ACD-SHIP technique

ACD-SHIP has two main modules; query extraction, content discovery and ranking. The meta-data in the CM model supplies query dimensions to the query extraction module to generate search terms. These terms are fed into a search engine and the returned results are input to the content discovery module which determines relevance of each search outcome by comparison with existing content on the same topic. The functionality of the two modules is discussed next.

Query Extraction

The query dimensions used in this module are primarily sourced from meta-data found in the first element of each term in the CM matrix. The element $a_x t_y$, denotes the audience sub-grouping and the term (or topic) from the domain ontology. The two terms alone do not suffice to generate a suitable search query. Two other dimensions, mandatory domain terms and synonyms for each term need to be introduced to the search queries. The mandatory domain terms introduce the specific health domain to the search query. Such terms limit the search from returning irrelevant results. For instance, if the SHIP is focusing on breast cancer, the term ‘breast cancer’ must appear in the search query. Other query dimensions such as the level of understanding of the audience sub-group (simple, advanced or scientific), publication date (latest, recent or old) can also be introduced to the search query.

The set of synonyms apply to all query dimensions discussed thus far. There could be synonyms for the audience sub-grouping, i.e. synonyms for ‘old women’ include ’50 years and above’, ‘mature women’ and so forth. There are also synonyms for the mandatory domain term and the ontology term such as scientific or medical terms (e.g. Lymphedema being the scientific term for swollen limbs).

The module will thus generate multiple search queries for each combination of $a_x t_y$ by making use of relevant synonyms for all three terms that compose the query. Each search query will be run separately on a search engine and the results merged into one distinct set after duplicates are removed. The final result, which maintains the ranking assigned by the search engine, is then input to the content discovery and ranking module.

Content Discovery and Ranking

This module initiates with a crawler that fetches all the search results and converts them to plain text. Search results with minimal content (such as index pages, announcements, discussion forums) are discarded leaving the resources with bulk textual content. The converter was developed using Apache Tika and thus able to parse most document formats, HTML, PDF and XML. The plain text of each resource is further normalised using stop-word removal and Porter’s stemming algorithm (Porter 1980) to represent a bag of words that captures the semantics of each resource.

A similar normalisation process is carried out on existing content that have been categorised using the same dimensions by the domain expert. The resources relevant for $a_x t_y$ are fetched from the content repository and merged into a single resource/document, which is also pre-processed using the same techniques above to produce a corresponding bag of words. This document represents the build-up knowledge within the repository as it attempts to capture domain expertise that is contained in the expert recommendations of content for the query dimensions of $a_x t_y$. It can now be used as a benchmark document to measure the relevance and quality of each new resource and also as a source for ranking. The vector space model (VSM) was used as the metric to measure relevance/similarity.

The VSM introduced by Salton et al in (Salton et al 1975) models documents as elements in term space. The VSM has been successfully applied to several text mining/business analytics applications such as ontology based information retrieval in (Castells et al 2007), incremental learning from text (De Silva and Alahakoon 2010) and disease identification (Sarkar 2012).

In VSM, each document is represented by a vector of terms in the document, and these vectors exist in term space, which is composed of all the unique terms in the collection. The normalisation process helps to generate a better representation of the document in term space. When comparing two such vectors, closeness is determined by measuring the angle between the two. It is typical to use weighted vectors to measure the similarity score and the TF-IDF (Salton and McGill 1986) computation is used as the basis to assign weights to the terms. This computation assigns a higher score to terms that occur with high frequency in a document (TF) relative to their occurrence in other documents in the collection (IDF). In comparison with the benchmark document, vectors (representing new content) with many high frequency terms would appear closer than vectors with less high frequency terms. Thereby, the relevance of new content to existing content can be identified and appropriately scored by applying the VSM on the normalised document sets obtained.

It is envision that the new content discovered by the proposed ACD technique will undergo review and approval of the domain expert as part of the content management lifecycle, thus contributing to the improved efficiency of the SHIP management process.

ACD-SHIP APPLICATION OUTCOMES

Current BCKOnline portal follows the content management lifecycle, which involves domain experts for the maintenance of the SHIP content. The layer of personalisation took several factors into account, including the type of audience, level of understanding, user roles, information preferences etc. (Burstein et al. 2005). A robust CM model was used by the domain experts to manage and revise the content. This CM model was instrumental in experimentation and validation of the ACD-SHIP technique proposed in this paper. The well-defined structure and interface to the CM repository contributed towards easy integration to the ACD technique. The aim of this study was to validate the usefulness of query extraction and text analytics technique to search for more relevant material based on the current content of the repository.

When applying the ACD-SHIP technique to the BCKOnline portal, several customisations were necessary. The technique is composed of two modules. For the first module, query extraction, it was possible to further sub-divide the audience sub-groupings and introduce multiples of these to the search query. The domain ontology as specified by domain experts was composed of 795 terms, including synonyms of certain terms. The synonyms were separated and further synonyms were found for the whole set of terms. The mandatory terms were limited to a few that strongly conveyed the illness as breast cancer. Figure 4 illustrates these components as customised for BCKOnline portal.

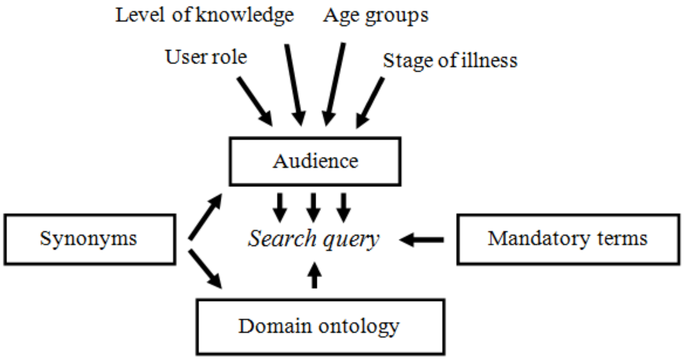


Figure 4: ACD-SHIP technique

Figure 5 presents the top 30 domain ontology terms in the content repository. The graph exhibits a long tail, where a larger number of the resources are categorised in smaller groups. This signifies the breadth of the health information for breast cancer and further justifies the need for an ACD process. The highest numbers of resources are on the primary sub-topics of early, advanced and recurrent breast cancer.

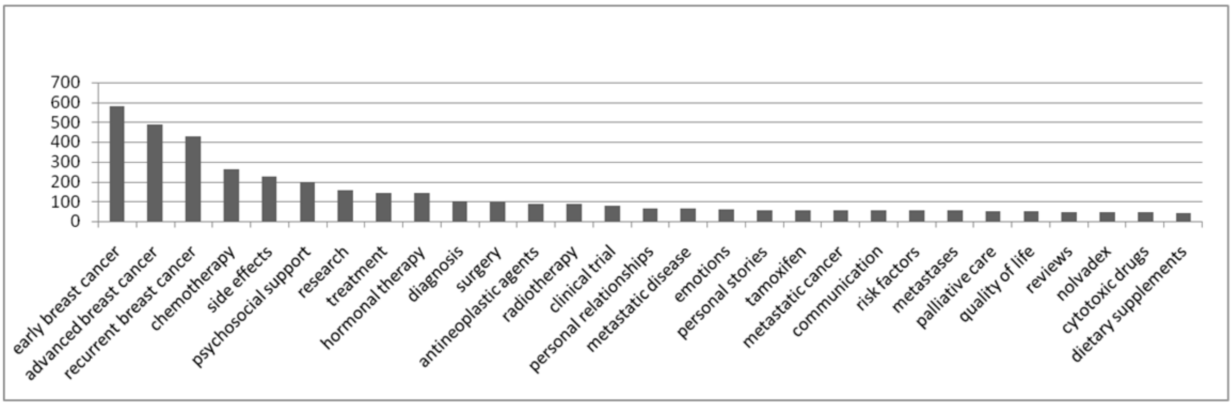


Figure 5: Top 30 domain ontology terms

Two terms were selected from this graph to demonstrate the ACD-SHIP technique. The terms are 'dietary supplements' with a content count of 43 and 'quality of life' with a count of 50. The first term was selected as it was the last in this top 30 distribution while the second, a general day to day term, was selected to demonstrate the robustness of the technique in picking up useful and relevant resources. The queries generated for these two terms are shown in Table 1.

Table 1: Output from query extraction module

Domain terms	**Search queries**
dietary supplements	'breast cancer dietary supplements' , 'breast cancer dietary supplements young patients', 'advanced breast carcinoma scientific nutrition', 'nutritional supplements recurrent breast cancer', 'nutritional supplements middle-aged patients breast cancer', 'early breast cancer dietary supplements'
quality of life	'breast cancer quality of life', 'recurrent breast cancer quality of life carers', 'advanced breast well-being middle-aged', 'recurrent breast cancer wellbeing old patients', 'breast cancer quality of life family'

Search queries were run on the Google search engine. The top 100 results for each search query were extracted, filtered and checked for duplicates. The lists of results were merged and input to the second module, content discovery and ranking.

In content discovery and ranking, each search result was downloaded to a local database and a plain text document extracted from the downloaded format. A VSM was generated from each document while separately a VSM was generated from the merged resources already in the content repository. After stemming and stop-word removal, the average term count of the downloaded documents was 2500 while the average term count of the merged content from the repository was 3000. Given the likeness of the mean term count of the document sets, it was possible to use Euclidean distance to compare semantics of new resources to the VSM of existing resources. The threshold for Euclidean distance between two resources was set to 0.75 in order to normalise the disparity in number of terms of the benchmark document with that of new individual resources.

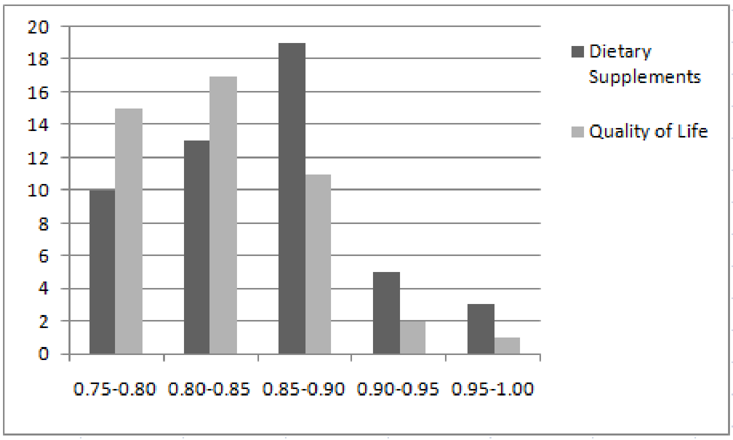
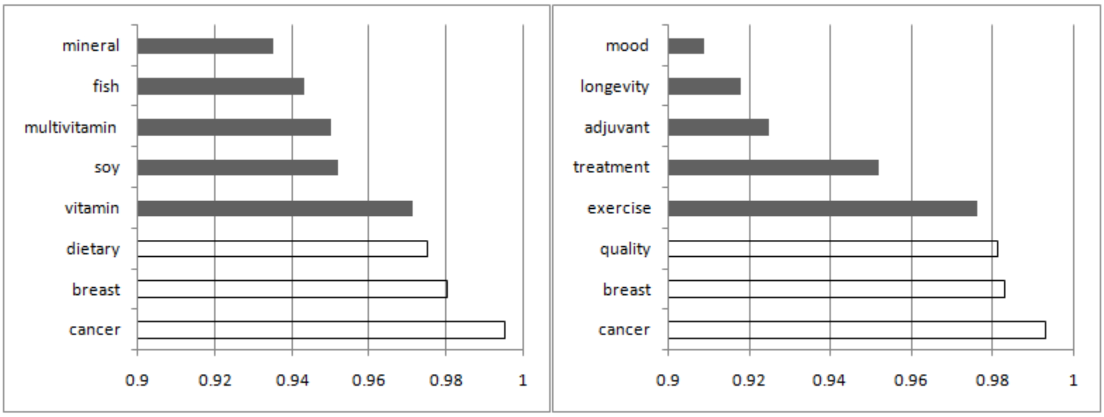


Figure 6: Histogram of closeness measures of new resources to benchmark VSM

For the first term, ‘dietary supplements’, there were 50 new resources above the threshold and 46 for the second term. Figure 6 depicts a histogram of the closeness measures. A majority of resources are between the 0.75-0.90 region with less above 0.95. This minor deficiency can be attributed to the large number of terms in the benchmark document in comparison to individual resources.

A further analysis was conducted on the text corpus to determine key terms in the new set of resources. The TF-IDF values for each set of documents was retrieved and summed per term. The top seven terms for each of the two searches is shown in Figure 7. The most frequent terms are domain specific terms; these have been greyed out in the graph. The remaining terms give insight to the content of the new documents and can also be used as input to the next iteration of the query extraction phase. For instance, the use of soy or fish as a dietary supplement could be a noteworthy exploration in the next iteration of content management.



(a) (b)

Figure 7: High frequency terms in new content (a) dietary supplements (b) quality of life

The proposed ACD-SHIP technique was successful in acquiring new and relevant content for the portal with approximately 45-50 resources ranking above 0.75 for the two selected terms. An additional outcome from the technique was the ability to identify high-frequency terms relevant to the search queries being processed. In order to further establish its usability, the following section reports validation of the technique based on existing SHIP resources.

ACD-SHIP Validation

The current portal resources were ideal to be used in validation as these have been meticulously examined for appropriateness by domain experts. The authors were unable to validate the query extraction module as the resources found in the content repository could not be easily located in the top results returned by search engines. This is mainly due to frequent updates to online content resulting from dynamic developments in the medical and health domains. However, the key semantics comparison task is carried out by the second module which could be readily validated using existing content. The content for all existing resources (‘dietary supplements’ - 43 and ‘quality of life’ - 50) were downloaded and a VSM generated for each. These VSMs were compared against the same benchmark document used above. The distance metric was based on Euclidean distance. Figure 8 plots the histogram for the closeness measures.

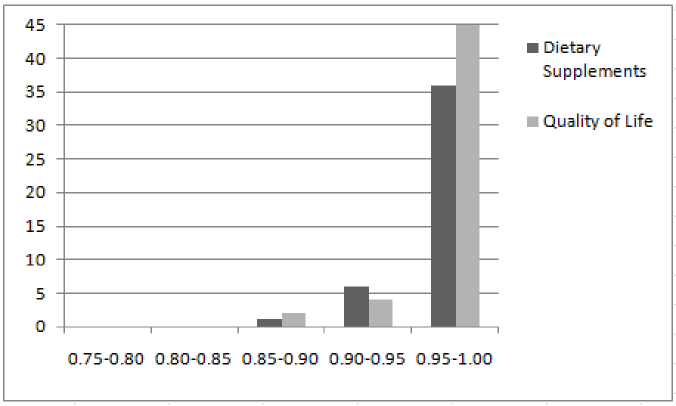


Figure 8: Validation- histogram of closeness measures of existing resources to benchmark VSM

The results were heavily right skewed with a large number of resources having high proximity to the benchmark document. The capability of ACD-SHIP to accurately compare semantics of two sets of text corpus is readily visible in this result.

The proposed ACD-SHIP technique is composed of query extraction, content discovery and ranking modules and fits well as a semi-automated support mechanism of an intelligent tool for the domain expert (Xie 2009). The first module generates queries from expert endorsed content that address several different query dimensions. These search queries are forwarded to the second module which downloads the actual content and compares it for relevance and accuracy with similar material in the content repository. The relevant content is ranked and forwarded to a domain expert who can then refresh the repository with up-to-date content. The ability to determine relevance and accuracy of new content returned by a search engine greatly improves the overall quality and timeliness of a HIP. The potential of such a tool to enhance the functionality of a SHIP is supported by encouraging results achieved when applied in the BCKOnline case as illustrated above. Furthermore, the ACD-SHIP technique needs to be manually evaluated by several experts across several domains. The outcomes from such an evaluation will be useful to improve the quality of content retrieved as well as the ranking method used.

DISCUSSION AND CONCLUSION

The core capabilities of a SHIP are content management, content delivery and collaboration. There are many intelligent technologies and techniques that can be adapted to enhance each of these capability, thus the emergence of SHIP, which is not only meeting requirements of the users in personalised health information, but is also capable to learn and adapt its content over time based on usage data, providing assistance with the content management task. This paper initially explored the content management capability of a SHIP in general and reflected on the experience of two multidisciplinary SHIPs development projects, which demonstrated a strong link between the quality of the content and the access to domain expertise for satisfying the needs of the users.

The paper formulated a generic CM lifecycle as well as a CM model for content acquisition. The CM lifecycle clearly emphasised the role of the domain expert in maintaining and updating the content and made a strong argument for supporting this effort as imperative for the long term viability of the SHIP. The proposed CM model explained the relationship between the stakeholders and information entities, and fused them into a single comprehensive structure indicative of the varieties of content and personalisation required.

Following this exploration into content management, the paper proposed and explained in details ACD-SHIP technique –an automated technique to acquire new and relevant content for maintaining the content repository. It is based on a text mining model, the VSM, and was tested on the BCKOnline SHIP with positive outcomes as reported in Section Four. Future research would require testing the usability of the proposed technique with the domain expert.

The authors are currently conducting research on integrating several other smart technologies that will enrich the core capabilities of a SHIP. These include semi-automated content summarisation, user annotations, intelligent feedback and recommendations. Content delivery can also be greatly improved with smart techniques such as user profiling, geo-location and interactive content. The successful adoption of SHIP as a point of reference for health related issues is a strong indicator of the expanding information needs of online communities. Therefore this study is timely and generates research outcomes that contribute towards development of comprehensive, state-of-the-art smart health information portals.

REFERENCES

Abbar, S., Bouzeghoub, M., Kostadinov, D., Lopes, S., Aghazaryan, A., and Betge-Brezetz, S. 2008. “A Personalized Access Model: Concepts and Services for Content Delivery Platform,” in: *Proceedings of 10th international Conference on Information Integration and Web-based Applications & Services (IIWAS'08), Linz, Austria*.

Aggarwal, C., and Zhai, C. 2012. *Mining Text Data*, Springer.

Bates, M. 2011. *Understanding Information Retrieval Systems: Management, Types, and Standards,* Auerbach Publications.

Boiko, B. 2001. *Content Management Bible*, Wiley Publishers.

Bonifacio, M., Franz, T., and Staab, S. 2008. “A Four-Layer Model for Information Technology Support of Knowledge Management,” in: *Knowledge Management: An Evolutionary View,* Editors: Irma Becerra-Fernandez, Dorothy E. Leidner, M E Sharpe Inc.

Burstein, F., Fisher, J. L., McKemmish, S. M., Manaszewicz, R., Malhotra, P. 2005. "User Centred Quality Health Information Provision: Benefits and Challenges," in: *Proceedings of the Thirty-Eighth Annual Hawaii International Conference on System Sciences, Los Alamitos CA USA.*

Burstein, F., McKemmish, S.M., Fisher, J.L., Manaszewicz, R., Malhotra, P. 2006. "A Role for Information Portals as Intelligent Decision Support Systems: Breast Cancer Knowledge Online Experience", in: *Intelligent Decision-making Support Systems: Foundations, Applications and Challenges*, Springer, London UK, pp. 359-383.

Castells, P., Fernandez, M., and Vallet, D. 2007. "An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval," *IEEE Transactions on Knowledge and Data Engineering* (19:2), pp 261-272.

Chau, M., Huang, Z., Qin, J., Zhou, Y., and Chen, H. 2006. "Building a Scientific Knowledge Web Portal: The Nanoport Experience," *Decision Support Systems* (42:2), November, pp 1216-1238

Ciccarese, P., Ocana, M., Castro, L.J.G., Das, S., and Clark, T. 2011. "An Open Annotation Ontology for Science On Web 3.0," *Journal of Biomedical Semantics*, (2:2), S4.

Collins, H. 2002. *Enterprise Knowledge Portals: Next Generation Portal Solutions for Dynamic Information Access, Better Decision making and maximum results*, American Management Assoc., Inc. New York, USA.

De Silva, D., and Alahakoon, D. 2010. "Incremental Knowledge Acquisition and Self-Learning from Text," in: *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2010), Barcelona, Spain.*

ECM3. 2009. "ECM3 Maturity Model," Creative Commons 2009-2010, Wipro-Real Story Group - Hartman.

Elmagarmid, A.K. and McIver, W.J., 2001. The ongoing march toward digital government. Computer 34, 32–38.

Erkan, G., and Radev, R. 2004. "Lexrank: Graph-Based Lexical Centrality as Salience in Text Summarization," *Journal of Artificial Intelligence Research*, (22:1), pp 457-479

Fisher, J., Burstein, F., Lazarenko, K., Lynch, K. and McKemmish, S. 2007. "Health Information Websites: Is the Health Consumer Being Well Served?," in: *Proceedings of Americas' Conference on Information Systems (AMCIS), Keystone, Colorado – USA.*

Fisher, J. L., Manaszewicz, R., McKemmish, S. M., Burstein, F., Malhotra, P., Moon, J., 2004. "User-Centric Portal Design for Quality Health Information Provision: Breast Cancer Knowledge Online," in: *Proceedings of the 15th Australasian Conference on Information Systems, Tasmania, Australia.*

Katz, R.N., 2002. Web portals and higher education: technologies to make IT personal. Jossey-Bass, San Francisco.

Kim, H., Ha, I., Lee, K., Jo, K., and El-Saddik, A. 2011. "Collaborative User Modeling for Enhanced Content Filtering in Recommender Systems," *Decision Support Systems* (51:4), pp 772-781.

Kukafka, R., Khan, S. A., Hutchinson, C., McFarlane, D. J., Li, J., Ancker, J. S. and Cohall, A. 2007. Digital partnerships for health: steps to develop a community-specific health portal aimed at promoting health and well-being, *AMIA Annual Symposium Proceedings*, vol. 2007, pp. 428–432.

Mandl, K.D., Kohane, I.S. and Brandt, A.M., 1998. Electronic patient-physician communication: problems and promise. Annals of Internal Medicine. 129, 495–500.

McKemmish, S.M., Manaszewicz, R., Burstein, F., Fisher, J.L. 2009. "Consumer Empowerment through Metadata-Based Information Quality Reporting: The Breast Cancer Knowledge Online Portal," *Journal Of The American Society For Information Science And Technology* (60:9), John Wiley & Sons Inc., Hoboken NJ USA, pp 1792-1807.

Middleton, S.E., Shadbolt, N.R., and De Roure, D.C. 2004. "Ontological User Profiling in Recommender Systems," *ACM Transactions on Information Systems*, (22:1), pp 54-88.

Pier, C., Shandley, K., Fisher, J.L., Burstein, F., Nelson, M., Piterman, L. 2008. "Identifying The Health And Mental Health Information Needs of People with Coronary Heart Disease, with and without Depression," *Medical Journal of Australia* (188:12), Australasian Medical Publishing Company, Pyrmont NSW Australia, pp. S142-S144.

Porter, M.F. 1980. "An Algorithm for Suffix Stripping", *Program* (14:3) pp 130-137.

Salton, G., and McGill, M.J. 1986. *Introduction to Modern Information Retrieval*, McGraw-Hill, Inc.

Salton, G., Wong, A., and Yang, C. S. 1975. “A Vector Space Model for Automatic Indexing,” *Communications of the ACM* (18:11), pp. 613-620.

Sarkar, I.N. 2012, “A Vector Space Model Approach to Identify Genetically Related Diseases,” *Journal of the American Medical Informatics Association* (19:1) pp 249-254.

Theofanos, M.F. and Mulligan, C., 2004. Empowering Patients Through Access to Information. *Information, Communication & Society* 7, 466–490.

Tatnall, A. 2005. *Portals, Portals Everywhere ...Web Portals: the New Gateways to Internet Information and Services*. Hershey, PA, Idea Group Publishing.

WCMM. 2009. *Web content maturity model*, Forrester Research, http://www.forrester.com/go?objectid=RES4656

Xie, J. and Burstein, F., 2011. Using Machine Learning to Support Resource Quality Assessment: An Adaptive Attribute-based Approach for Health Information Portals, *Proceedings of the 16th International Conference on Database Systems for Advanced Applications International Workshops*, pp. 526-537

Xie, J. 2009 “Sustaining Quality Assessment Processes in User-Centred Health Information Portals,” in: *Proceedings of the Americas Conference for Information Systems, AMCIS 2009.*

ACKNOWLEDGEMENTS

This research is partly supported by the University of Ballarat 'Self-sustaining Regions Research and Innovation Initiative', an Australian Government Collaborative Research Network (CRN) and the Telematics Trust, project: Breast Cancer Knowledge Online (BCKOnline) Portal.

COPYRIGHT