Adaptive Knowledge Retrieval Using Semantically Enriched Folksonomies

Applications in the domain of homemade explosives

Dimitrios Pappas, Iraklis Paraskakis
South-East European Research Centre (SEERC)
The University of Sheffield International Faculty, CITY College
Thessaloniki, Greece
{dimpappas, iparaskakis}@seerc.org

Abstract—Organizations invest considerable development effort in personalized systems to reap the benefits of knowledge management. Competition and the need for innovation imply in ever-changing and emerging needs that result in costly effort to redesign such systems. A flexible knowledge management approach should achieve personalization for organizations by adapting to the domain of the organization, meanwhile simplifying the incorporation of future knowledge retrieval needs. We describe such an approach that employs a knowledge management platform at its core, which supports adding and removing multiple folksonomies, constantly enriches them with semantics and provides domain-specific recommendations and semantic search capabilities. This approach to knowledge retrieval has been applied to the domain of homemade explosives and counter-terrorism efforts as part of the HOMER project. Preliminary evaluation from the perspective of the end-users – law enforcement, security and related agencies – is presented.

Keywords—folksonomy, personalization, adaptive knowledge retrieval, recommender systems, semantic enrichment, result precision

I. INTRODUCTION

The process of Knowledge Management (KM) revolves around providing efficient and effective means of recording, cultivating, sharing – internally or externally – and capitalizing on organizational knowledge. KM efforts are being motivated by the potential to leverage organizational learning, raise expertise across organization employees, increase network connectivity and fostering innovation, among numerous other benefits. Knowledge retrieval, a highly valued aspect of the KM process, seeks to return information in a structured form compliant with the human mental process of knowledge acquisition and understanding. Unlike the more basic processes of data and information retrieval, knowledge retrieval is heavily burdened with the need to achieve proper representation and accurate extraction of knowledge. Consequently, it tends to shy away from deterministic or statistical models, artificial or natural language-only queries and representation by numbers, rules and markup language; instead, it focuses on semantic and inference models, access via natural language complemented by knowledge structure and representation by an assortment of complex approaches such as ontologies, semantic networks, predicate logic and concept graphs [1].

A key factor for the success of the knowledge retrieval process is its capability to produce relevant results to the users amidst the information overload they are typically presented with. To this end, any organizational KM effort should account for the needs of the users and, by extent, to organizational objectives, a process which is referred to as personalization and it is becoming increasingly central to successful knowledge retrieval [2]. Personalization methods vary greatly and are usually combined to achieve optimal results, ranging from the analysis of user behavior (e.g., previous searches, frequently accessed content) and context (e.g., profile, location) to exploitation of the knowledge representation methods (e.g., semantics, ontologies, taxonomies) [3]. The most popular mechanisms employed for the purposes of knowledge retrieval nowadays are the recommender systems, which employ the aforementioned methods in order to suggest more relevant resources to the user during their interactions with the platform (e.g., search, browsing) [4, 5].

Recommender systems generally fall into one of the following three categories based on which type of information available to the KM system they operate on: (a) content-based filtering, (b) collaborative filtering and (c) hybrids combining elements of both (a) and (b) [6]. Content-based filtering aims to study the characteristics of a resource and compare it to those of other resources in order to recommend items that possess similar properties. Collaborative filtering, on the other, hand focuses on studying user behavior, whether it is done explicitly (e.g., search, rating, favorites) or implicitly (e.g., resources accessed, preferences of user social network), in an attempt to predict resources that users would be interested in.
Personalizing knowledge retrieval with the use of recommender systems has seen an explosive growth in recent years, especially with web-based stores and social networking [4]. Related work in this field revolves around achieving better personalization by focusing on fine-tuning analysis and prediction algorithms, as well as the development of hybrid systems that manage to tackle limitations imposed by separate content-based or collaborative filtering systems [3, 7]. As such, these attempts often disregard important factors such as the inconsistency of human information retrieval behavior and its relation to the domain of the user [8].

We take this opportunity to propose an approach that aims to bring everything together: a Knowledge Management Platform (KMP) based on the combination of folksonomies, semantic networks and user behavior analysis that can effortlessly be parametrized to meet organizational needs. It offers intuitive and multi-faceted knowledge retrieval that manages to adapt to the domain of the organization and aspires to tackle the problems of user inconsistency and result precision. This is achieved primarily with the utilization of semantic search and recommender systems.

The three major steps to achieving it are the following:

1. Implement a number of folksonomies specific to the domain of the organization.
2. Protect the folksonomies and refine them with semantics towards knowledge organization.
3. Reap the benefits of semantically enriched search and domain-specific recommendations.

The capacity for adaptation is showcased with the application of this approach to the Homemade Explosives (HME) domain, with the objective of assisting counter-terrorism efforts (e.g., investigations, threat response) as an integral part of the homemade explosives characterization and capability research project, HOMER. The simplified process of creating the specialized HOMER KMP with minimal development effort is discussed. Preliminary evaluation results indicate that it succeeds in offering relevant recommendations to users as part of their investigative activities. Furthermore, the implementation of semantic search features in the KMP appears to be a promising starting point for narrowing the gap between searching and the inconsistency of human information retrieval behavior.

II. ADAPTIVE KNOWLEDGE RETRIEVAL

A. Conceptual Architecture

The conceptual architecture of the proposed KMP in Fig. 1 illustrates that the core of the system resides in the knowledge base, which facilitates the smart processing of incoming information through indexing mechanisms (text analysis, tags, taxonomies components), and which is exploited by the knowledge retrieval mechanisms and recommender systems provided to the users (semantics, browsing history, tags, browsing content, semantic search components).

The users have limited access to the system which is governed by their role in their organization (role-based access component). Apart from knowledge retrieval, users naturally participate in knowledge recording and this is performed via the content generation components (blogs, Wikis). All these components together realize the four important aspects of KM: creating, developing, sharing and using organizational knowledge.

Employees of an organization are assigned specific roles that come with varied rights and responsibilities. This particular mechanism ensures that the KMP can leverage the actualization of an organizational structure: only certain users may create knowledge and users may selectively consume knowledge as the organizational policies allow for it.

Intuitive Web 2.0 technologies for capturing, sharing and developing knowledge, such as Blogs and Wikis, enable the option of categorically organizing resources (e.g., exports reports blog, imports reports blog), as well as structuring the information hierarchically if needed (e.g. department descriptions Wiki). Each entry is in the form of a typical document enriched with file attachments and supporting rich-text and HTML content, with the additional capability of directly referencing other related documents if the author deems it necessary.

B. Folksonomy-based Knowledge Management

As part of the content input process, the users are obliged to go through the process of tagging their documents. There is no predefined, locked taxonomy of tags in the KMP, instead this process is based solely on the utilization of folksonomies. As a consequence, users have considerable control over the organization of information, a definite benefit for personalized knowledge retrieval. This stems from the fact that the tagging process encourages users to think about the domain, how they and other users would need to access this information, hence it reflects their way of thinking about the domain.

Collaborative tagging systems based on folksonomies are adaptive when it comes down to organizing information and could perform better than orderly classification systems [9]. Another inherent property of such systems is the option to investigate the frequencies of the tags being used and derive valuable statistics about the domain, the knowledge within the system and emerging trends [10].
While tagging solutions have been praised for several aspects related to knowledge management, they have also received considerable criticism regarding their effectiveness. Researchers primarily attribute these issues to the potential ambiguity of tag meanings and the possibility of overlapping tags due to synonyms, as well as to the lack of a centralized, controlled vocabulary [9]. These arguments state that when such issues take root within a folksonomy, it might be impossible for the users to develop a consensus and for a common vocabulary to emerge, therefore knowledge organization and retrieval are crippled.

It can be argued that language-related problems are a by-product of an extremely large and diverse user base on a general-purpose application, whereas domain-specific applications expected at organizational level are bound to follow a much stricter and narrow vocabulary that users may employ. Domain and content-specific tag recommendations provided to users during the process of content creation also contributes to alleviating such issues [9]. Furthermore, empirical studies have refuted the issue of consensus by indicating that, over time and with increasing numbers of recorded resources, folksonomies tend to turn into stable distributions despite the absence of a centralized and controlled vocabulary; in some cases common structures at a categorization level can also be developed [11].

The benefits of folksonomies to knowledge retrieval on an organizational level prompt us to incorporate this solution as the basic tool for knowledge management on the KMP, while taking under consideration the potential pitfalls it is bundled with. To this end, we have employed several prevention and counteracting mechanisms that possess the potential to convert the folksonomy approach into a robust solution for knowledge organization and development, in addition to retrieval.

1) Multiple, domain-specific folksonomies

By studying the domain and breaking it down to a condensed set of sub-domains, we are able to narrow down the vocabulary that users may employ to tag resources for each one. Naturally, this enables other KM methods we employ (semantic enrichment, recommendation systems) to capitalize on the variety of vocabularies being formed and produce fine-grained knowledge retrieval processes, as well as to achieve multi-dimensional knowledge representation.

2) Folksonomy-based tag recommendations

Introducing multiple folksonomies that users are obliged to employ during content generation, provided that the content they are generating applies to all of them, may inadvertently discourage them from using the knowledge recording process to its fullest extent. We counter that with tag recommendations based on a hybrid recommender system using semantics and Natural Language Processing (NLP) techniques. A welcome by-product of this approach is that it has the added benefit of tackling issues regarding confusing tag meanings and synonyms.

3) Content validation

Erroneous or inappropriate tagging by a single user may contaminate the folksonomy and, as a consequence, the semantics and recommender systems. This would result in contamination of the overall knowledge residing in the platform; the outcome could be in the range of insignificant to disastrous, depending on the domain and applications of the system. In order to minimize that risk, a content validation mechanism has been implemented. Expert users review newly created content and either approve it or request changes from the author. The resource is still accessible by all users with access rights and they are warned of its validations status, however it is not incorporated into existing knowledge unless it has been validated; i.e. it does not appear as part of the semantic search and recommendations.

4) Folksonomy refinement

As part of the semantic enrichment of the folksonomies in this proposed approach, one technique that has been employed is the refinement of the folksonomy structure. This mechanism essentially attempts to give a structure to the folksonomy, as the consensus on the vocabulary is getting stronger by consulting an external knowledgebase. It enhances representation and retrieval of knowledge through the recommender systems.

C. Semantic Enrichment of Folksonomies

Flat, unstructured and unsupervised vocabularies in folksonomy-based KM systems have been found lacking in result precision due to the lack of hierarchical organization of knowledge and contamination with ambiguous or irrelevant tags [12]. This endeavor is a popular practice within Web 2.0 systems, and on the way to realizing the semantic web, because of the core benefits that semantics have been proven to bring to folksonomy systems: structure and its protection, as well as search enrichment, all of which indirectly bring higher search and navigation precision [12].

Semantics broadly refers to the study of meaning in language; in the case of tags and the field of knowledge management, it is the study of the relationship between the word and its denotation. By clarifying this relationship of a tag and its meaning within the current context, identifying a concept, as well as tag interrelationships, structure and hierarchy may emerge; consequently, precision, navigation and representation are enhanced. The techniques employed attempt to formulate semantic models, such as semantic networks, to ultimately generate a conceptual view of the data that is a close representation of the real world. The relations identified among concepts vary depending on the objective of the methods followed, usually focusing on finding equivalent terms, subsumption relationships, hierarchical higher and lower-level concepts, to name a few [12].

Folksonomy enrichment with semantics tends to rely on three major contributing sources [12]:

- Folks (i.e. the users, their process & motivations).
- External knowledge-based sources (e.g., DBpedia and Wikipedia, Ontologies).
- Statistical and mathematical techniques.

The proposed KM system includes components in its architecture, as outlined in Fig. 1, that exploit these sources. To begin with, the investigation of sub-domains for the creation of
multiple domain-specific taxonomies, in conjunction with the expected role-based access of the users, focuses on the “Folks” source. Understanding the expected user behavior and their actions within organizational context is the first and most invaluable step. Ultimately, by understanding how users respond to tagging and by observing how they access information, the proper partitioning of the domain emerges. Once the new folksonomies are populated with terms and subsequently enriched, the resulting recommendation and search precision increase becomes apparent through the underlying semantic network formed for each one.

External knowledge-based sources, specifically DBpedia, as well as statistical techniques, are utilized by both the tag recommendation component and the text analysis indexing mechanism. Tag recommendation uses a tag extraction function based on statistical NLP techniques. It attempts to identify key terms within the provided content and consults DBpedia, using the concepts category and upper or subcategories in the external source that corresponds to the sub-domain (e.g. DBpedia Category: Countries and Cities, folksonomy and sub-domain “locations”). The algorithm assigns a rating to each identified term and returns a list of suggestions with those possessing high ratings, therefore assisting the users with the process of tagging while protecting taxonomy structure, too.

The text analysis indexing mechanism is the most complex sub-system and leverages the folksonomy refinement technique. Essentially, it performs similar functions as the tag extraction but for different purposes. The NLP component compares knowledge artifacts with each other in order to come up with statistics on which artifact is semantically similar to another. Through exploitation of DBpedia information, ratings are assigned to tags again, but this time it goes further and tag semantics are compared with each other to try and find the semantic connections among all of them. Ratings are assigned to each established connection, relationships are identified (e.g. parent-child concepts) and for each concept related to another the common semantic links are stored, too. Eventually, a fully-fledged semantic network of the domain is formed, while this refinement method manages to construct an invisible hierarchy for each folksonomy, too.

The taxonomy refinement and tag extraction NLP mechanisms have been borrowed in part from the OrganiK project, which proposed a socio-technical approach to KM [13]. They have been heavily modified for this project and the external source support for DBpedia has been added.

D. Recommender Systems

The final link in the chain of techniques and mechanisms towards adaptive knowledge retrieval comes in the form of the recommender systems. These recommender systems complement the basic means of retrieving knowledge in the proposed architecture: perform full-text search, browse content categorically via each different blog or Wiki, browse content by a single tag it was annotated with, browse content from latest activity lists and browse content by author.

In the approach that we follow here, the recommender engine of the KMP is tightly coupled with the semantics engine; ratings and links of the semantic network behind the tags and folksonomies are exploited to provide accurate recommendations in various forms. It is imperative to stress that, while all of the recommender systems described below are enabled by default, an organization deploying the KMP may disable any number of them to achieve optimal results.

1) Browsing history

This is a typical recommendation system with a twofold purpose: provide for each user a list of resources similar to those they tend to access, and for each resource currently being accessed to provide a list of resources that other users who accessed this one, have accessed, too.

2) Tag similarity

Another typical recommendation system that does not make use of the semantic refinement of the folksonomies. It attempts to identify items that share tags with the one currently being accessed and provides a list with all of them to the user. It does not distinguish between different folksonomies, merely focuses on number of similar tags instead.

3) Content similarity

This system entirely forgoes the tag annotations and focuses solely on similarity of the content itself. NLP methods attempt to find linguistic similarities within two items and a ranking is assigned to the pair. Highest ranked pairs where the currently accessed item is part of are displayed as recommendations to the user.

4) Tag recommendations

Put to action during the content generation process in the KMP, this system primarily supplies the user with a set of tags for each folksonomy based on the text they have provided to simplify and expedite the process; to a lesser extent, it protects a folksonomy from potential contamination. It is context-aware in that it restricts identified terms for each folksonomy only to those that are semantically relevant to it. Users may either disregard the suggested tags or complement them with their own as they see fit, in which case the system recommends tags to them as they start typing. This may prevent tag duplication (e.g. existing term in lowercase, new term in first word uppercase) in the event that the NLP component recommender system failed to extract a term as a potential tag.

5) Domain-specific recommendations

The first of the two major contributions to knowledge retrieval personalization for organizations, this recommender system attempts to offer the resources that are most semantically relevant to the one currently being accessed. The chosen resources are displayed in separate blocks, one for each chosen folksonomy.

Semantically relevant resources in this particular case are identified as those that primarily have the most common links with the currently accessed one in the underlying semantic network. In case of similar number of links, resources with the highest ratings assigned to those links are preferred instead. The links, but not the ratings, are designed to be visible to a user in the KMP if they wish to study them and understand why they received these recommendations. Once a folksonomy has been enriched with sufficient terms and linked content, the semantic refinement process may facilitate stability through
structure and the emergence of a consensus. Effectively, this may lead to users discovering information, which they could not have known that it is in fact relevant to them.

6) Semantic search

The second crucial aspect of achieving personalized knowledge retrieval in an organizational domain comes in the form of a semantically enriched search function. Without the text analysis indexing mechanism and folksonomy refined, this would serve as a simple exact or best match full-text search. With the semantic enrichment, however, it is possible for the search function to attempt and expand search results with resources that do no match entirely the specified term(s), but have been deemed semantically relevant. The semantically similar terms are identifiable much like an exact match would be (e.g. bold font), in order for the user to understand why they are receiving this result. As another precaution to avoid confusion, these results follow after the exact and best matches.

Another property of the semantic search is that it accompanies the free-text search with recommendations that are, once more, categorized by folksonomy. The differentiation from the domain-specific recommendations is that, in this case, the recommended items are not semantically relevant resources, rather they are semantically related concepts (i.e. tags). Clicking to enable on one or many of them filters the search results to display only those that the selected concept is semantically relevant to. Content need not necessarily be tagged with the concept, just be related to it from another tag semantically linked with the chosen one.

The benefits of this semantic search and result filtering approach can be immense. Users need not attempt to come up with the one, perfect query to get the result they need; they can start from something they believe is relevant to it and, through an intuitive and simple process akin to the human investigatory behavior, finally arrive at their destination. This approach also provides counter-measures to alleviate cases of information overload where the user could be overwhelmed with the amount of information they would have to go through. Finally, an added benefit is that the semantic relation depth of the recommended concepts can be specified to the desired number of higher or lower level concepts that may be included; extremely fine-grained knowledge retrieval can thus be achieved with all these properties.

III. APPLICATIONS IN THE HME DOMAIN

A. The HOMER Knowledge Management Platform

The HOMER project aspires to bring together researchers and industry experts from the field of HME, as well as security and Law Enforcement Agencies (LEAs) related to the domain, with the aim of bolstering counter-terrorism activities and mitigating the threat of HME from the criminal and terrorist elements. This research project aims to enable maintainable, substantiated HME knowledge, which has been mapped, researched, collected, studied, filtered, analyzed and tested. Subsequently, the knowledge must be made available for LEAs, security and the manufacturers of precursor chemicals, in a secure and usable manner, and promote collaboration among the involved parties. The HOMER KMP is the enabling system, as far as knowledge-related activities are concerned.

The HOMER KMP has been developed over the Drupal content management system, which already provides mechanisms realizing some typical recommender systems (e.g. browsing history), but also offers an Application Programming Interface (API) that can be used to connect custom-made recommender systems and powerful, industry-standard search engines, in this case Apache Solr. The core effort towards adaptive knowledge retrieval came in the 3-step approach we propose: (1) determine the domain specific folksonomies, (2) configure their enrichment with semantics and (3) enable the proper recommender systems.

The 1st step was an amalgamation of work within the project consortium indicating what end-users may be looking for and what are the major areas of the HME domain. The former indicated that the available recommender systems and knowledge retrieval mechanisms are sufficient, while the latter resulted in the creation of three different folksonomies implemented as Drupal vocabularies: Precursors, Construction and Incidents. The process of adding vocabularies to the KMP is as trivial as clicking a button in the administration menu, making the addition or removal of folksonomies to adapt to future organizational needs effortless regarding this step.

The 2nd step of our process dictates the semantic enrichment of these folksonomies and this is where development effort is required. Specifically, the functions that perform tag extraction and folksonomy refinement have to be edited to point at the corresponding DBpedia categories being used to derive the semantic links. In this case, when the Precursors vocabulary is encountered the DBpedia category consulted is that of Chemical Components and its subcategories, for Construction it is Materials and for Incidents those of People, Places and Organizations.

The 3rd and final step is the decision on which of the recommender systems will be enabled and how much fine-tuning is required. This fine-tuning is accessible from the administrative interface through the form of weights on certain aspects of each knowledge artifact (e.g., title, body, blog-category it belongs to), or which of these aspects to actually include in the text analysis process. This is an iterative process that requires sufficient resources on the platform and extensive feedback from the end-users to ensure that the KMP. As a starting point for the HOMER KMP we kept the default proposed settings, which are meant to be put to the test during the two test cycles by the end-users to determine whether further parametrization is needed or they can accurately predict relevant resources. In the case of HOMER, internal evaluation without the input of end-users cannot accurately convey results due to the lack of availability in resources; almost the entirety of the material regarding HME is considered classified by LEAs and, consequently, inaccessible without authorization.

B. Preliminary Evaluation Results

The HOMER KMP has already undergone a test cycle and it is currently in its second, which has a larger focus on knowledge retrieval capabilities. Results of the first test cycle mostly focused on functionality and usability of the system,
however they included brief end-user feedback regarding knowledge retrieval capabilities. Preliminary results from the ongoing second test cycle regard result precision. The end-users involved in both cycle are LEAs highly active in the field of HME and counter-terrorism: the Police Service of Northern Ireland (PSNI), Guardia Civil (GUCI) in Spain and the Bulgarian Defense Institute (BDI).

End-users remarked that browsing history was not as convenient for them as the bookmarking of resources, due to their need to focus on specific resources (e.g. IRA, Dissident Republicans for PSNI) and the unwanted noise generated by browsing recommendations due to browsing other interesting resources available on the KMP. Content and tag similarity recommendations were found to be relevant but not frequently used, because the users expressed their preference for a more structured approach. Regarding this finding, the domain-specific recommendations were of great interest to end-users.

During the ongoing second test cycle approximately 200 HME recipes have been added to the KMP, contrary to a small number of incidents that LEAs may submit, but the users expressed that there is a high relevancy in the recommendations. Specifically, chemist users remarked that recipes and explosives with highly similar properties have been pushed to the top of the recommendation list in the Precursors folksonomy properly, while LEA users observed recommendations in the Precursors and Incidents folksonomy that are, in fact, tied to the report that they have been reading and to that particular organization involved. In some cases, users commented that they happened to get recommendations relationships that could be obscure to them personally (e.g. a junior investigating officer discovering a link between Terrorist Organization A and a specific type of electronic detonator through from the Construction and Incidents folksonomies).

The semantic search feature and result filtering through recommended concepts received a bulk of comments regarding usability and relevance. While the users found the process somewhat alien at the beginning, explaining that in LEAs they rely on typical database querying systems with pre-defined queries and tabular input forms, it proved to be intuitive and effective at reducing information overload with the added filtering options. In fact, the process of narrowing down the results with these terms appeared to comply with their investigatory cognitive process. As an example, users praised the ability to search for precursors and, using the recommended terms by the semantic search, locate what types of IEDs have been discovered, then where those IEDs have been used in the past and to ultimately end up with list of relevant organizations.

IV. CONCLUSIONS

We proposed an approach to achieving personalization for organizations through adaptive knowledge retrieval. Key methods employed are the domain-specific folksonomies, their enrichment with semantics and their protection from common pitfalls of the social tagging approach. The approach has received positive results according to preliminary evaluations, which include intuitive knowledge retrieval process consistent with human investigatory behavior, high relevancy of domain-specific recommendations and a quick tagging process.

The application of this approach to the intricate domain of HMEs has shown promise with regards to adaptive knowledge retrieval and outlines a simple, 3-step process that requires minimal effort to add or remove contextual dimensions as organizational needs change. Nonetheless, results from more in-depth evaluations (second test cycle) are required to determine the potential of the current configuration of recommender systems and identify any needs for fine-tuning.

Future work revolves around obtaining unclassified content by the end-users to conduct internal tests, complementing those currently being carried out by end-users. Finally, an attempt to bring all three of the steps of this process to the level of a KMP administrative interface is being considered (currently 2nd step is missing), therefore providing a framework that does not require code development to achieve organizational goals.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union’s FP7 Programme-Capability under grant agreement No 312883.

REFERENCES
