Exploiting Semantic Annotation of Content with Linked Open Data (LoD) to Improve Searching Performance in Web Repositories of Multi-disciplinary Research Data

Arshad Khan[1(✉)], Thanassis Tiropanis[1], and David Martin[2]

[1] ECS, University of Southampton, Highfield, Southampton, UK
{aaklvll, tt2}@ecs.soton.ac.uk
[2] Geography and Environment, University of Southampton, Southampton, UK
d.j.martin@soton.ac.uk

Abstract. Searching for relevant information in multi-disciplinary repositories of scientific research data is becoming a challenge for research communities such as the Social Sciences. Researchers use the available keywords-based online search, which often fall short of meeting the desired search results given the known issues of content heterogeneity, volume of data and terminological obsolescence. This leads to a number of problems including insufficient content exposure, unsatisfied researchers and above all dwindling confidence in such repositories of invaluable knowledge. In this paper, we explore the appropriateness of alternative searching based on Linked Open Data (LoD)-based semantic annotation and indexing in online repositories such as the ReStore repository (ReStore repository is an online service hosting and maintaining web resources containing data about multidisciplinary research in Social Sciences. Available at http://www.restore.ac.uk.). We explore websites content annotations using LoD to generate contemporary semantic annotations. We investigate if we can improve accuracy and relevance in search results affected by concepts and terms obsolescence in repositories of scientific content.

1 Introduction

Current searching techniques in web repositories[1] are predominantly based on keyword instances which are matched against content paying almost no attention to analyzing semantic meaning, types of content, context and relationship of keywords and phrases. Users of such web repositories have to rely on the incidental mention of the keywords and phrases in web pages, which is a challenge for users due to information overload of today's digital age. This issue is further complicated when we look at it across various disciplines where change in language terminology and concepts might change the meanings of today's web resources and other web-based content.

[1] A web repository stores and provides long term online access to a collection of web sites or web resources (containing static and dynamic web pages), research papers, presentations, experimental code scripts, reports etc. funded by UK research councils. Examples include http://www.data-archive.ac.uk/, http://www.timescapes.leeds.ac.uk/, http://www.icpsr.umich.edu/icpsrweb/ICPSR/.

P. Braslavski et al. (Eds.): RuSSIR 2015, CCIS 573, pp. 130–145, 2016.
DOI: 10.1007/978-3-319-41718-9_7

According to [1], search engines have experienced impressive enhancements in the last decade but information searching still relies on keywords-based searching which falls short of meeting users' needs due to insufficient content meaning. Similarly [2] terms the basic Web search as inadequate when it comes to finding contextually relevant information in web archives or collection of web sites like the ReStore[2] repository. Relationship between content must be an essential component to search results retrieval in such repositories but it is often missing due to the full-text keywords-based searching.

Figure 1 shows a typical web resource development and archival process involving funding bodies e.g. UK Research Councils, multi-disciplinary teams of researchers, higher education institutions and publication of research outputs in a dedicated online space either provided by the hosting institution or websites hosting company. The users of such research resources (according to our website survey in 2011 and 2013) are predominantly research students and fellows, academics, industry professionals and even funding bodies. Figure 1 outlines the entire process starting from a funding body funding a project in a higher education institution; particular research groups work on the project (typically for 3–6 years) and publish research outputs (Web resource in the form of a website) usually on hosting institution's website. These web resources end up later on in a Web repository to sustain its content for long term online access.

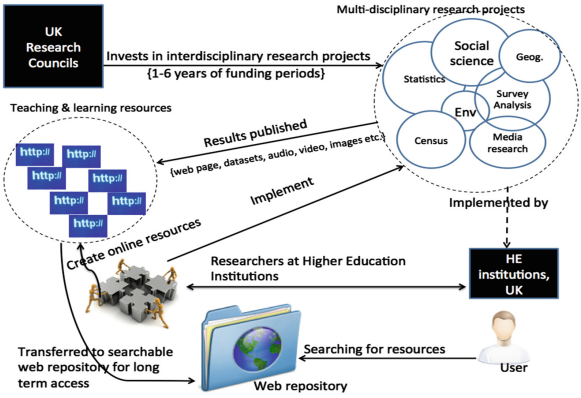

Fig. 1. An overview of web resource, creation, development and its archival into a web resource repository e.g. ReStore repository [3]

The inability to designate unambiguously the rapidly growing number of new concepts generated by the growth of knowledge and research in Social Sciences [4] is another issue failing the traditional search engines. Such issues have partly been addressed by keywords based searching where plain keyword queries are converted into equivalent semantic queries followed by syntactic normalization, word sense disambiguation [5] and noise reduction. To do that, the use of dictionaries (e.g. Wordnet), thesaurus and other library classification systems have been exploited in collaboration

[2] ReStore is an online repository of web resources developed as part of Economic & Social Research (ESRC) council funding-available at http://www.restore.ac.uk.

with the domain specific ontology to express keywords in more structured language. The semantic keywords are then matched with ontology terms and various semantic agents are applied to disambiguate terms and words before retrieving the results [6].

However, as described above, like other information domains, in scientific research disciplines terms change with the passage of time due to various factors e.g. cultural, social, technological, scientific and socio-economic etc. which compromise accuracy in search results. All of this suggests that semantic expressions and matching terms with ontologies classes/properties (linguistic) and instance data (semantic information) (in unstructured and heterogeneous content of repository websites) will not survive for too long and would need frequent and regular human intervention.

To address these issues, we have been focusing on 3 main areas as part of this research to see if we can improve the performance of online search applications widely used by users, who in our case, are researchers of a particular field such as the Social Sciences. These areas include (1) whether obsolescence in terms and concepts in online repositories of social science could be addressed by aiding keywords index with semantic annotation for better searching?, (2) whether a shift from domain-specific ontology-based annotation to distributed and wider data spaces like linked open-data (LoD)–based semantic annotation could address the issue of entity, concepts and relations disambiguation and thus result in better search results (high precision without compromising recall) and (3) whether crowd tagging (assigning semantic tags to relevant search results) could be employed to address the issue of content heterogeneity and obsolescence thereby benefiting the research community. This paper has so far investigated the first two points i.e. 1, 2 and thus our findings will only focus on the first two points in the rest of this paper. We will cover point 3 in the next phase of experimentation and evaluation.

This paper contributes to the related research by exploring LoD-based semantic annotation of heterogeneous websites content (ReStore repository in this case); scalability of annotation, indexing and retrieval when dealing with several content types; named entity-based linking of content based on semantic expressivity of users' keywords; and evaluating search results retrieved from the resultant semantic index, based on users' approval tagging (of system generated NE tags) and individual result ranking. Our approach further entails (a) development of an annotation, indexing and searching framework for contemporary searching in web repositories of scientific data; (b) automatic and manual annotation of content in the ReStore repository; (c) the deployment of a purpose built search engine called multi-faceted SocSci Search engine[3] for evaluating search results. We also discuss the technical deployment for various annotations and indexing approaches in the context of ReStore and NCRM[4] repositories and consider appropriate evaluation benchmarks at the time of searching and evaluation.

To evaluate search results, we have developed a custom-built search application sitting on top of full-text content, topical keywords, concepts and entities extracted from the selected documents corpus. We have conducted two distinct experiments i.e.

[3] Elasticsearch is a flexible and powerful open source, distributed, real-time search and analytics engine. Available at http://www.elasticsearchorg.

[4] http://www.ncrm.ac.uk.

searching performed by expert evaluators using full-text (keywords, content, title) searching technique and expert evaluation based on semantic indexing. In Sect. 2, we will review relevant work done in this area. In Sect. 3, we will elaborate the process of content annotation in repositories of heterogeneous content. Section 4 will explain the implementation of the indexing and retrieval framework and Sect. 5 will present evaluation results and lesson learnt. Section 6 will detail future work and conclusion.

2 Related Work

We recognize that some substantial work has already been done where the emphasis has been on collecting; storing and maintaining individual web resources in multi-disciplinary web repositories. However, searching across research repositories remains an open challenge. [1] highlights the limitations of keywords-based models and proposes Ontology-based information retrieval by capitalizing on Semantic Web (SW). However, poor usability of the systems usable by potential users and in completeness when applying search to heterogeneous sources of data still remain an issue.

The primary goal of any searching or retrieval system is to structure information so that it is useful to people while they search for information effectively and efficiently. Ontology-based semantic metadata extraction and storage have been around [7] since more than a decade but given that designing and evolving domain-specific ontologies still remains a challenge, alternative approaches have been adopted to extract relevant and meaningful information from text. For instance semantic indexing and retrieving the resulting knowledge base in a scalable manner still remains an issue. The semantic web research community has been experimenting with a fixed set of documents corpora with non-user friendly web based interfaces, which limit users' browsing capacity to visualize search results thus affecting overall information utility. Another major problem, the semantic web community faces for the construction of innovative and knowledge-based web services is to reduce the programming effort while keeping the web preparation task as small as possible [8].

We have seen user's query expansion-based searching and ontology-based information retrieval model proposed by [1, 9] and ontology-extension model based on adding further classes to the root ontology in [10] and ontology classes/properties matching between LoD cloud datasets (DBPedia, Freebase, Factforge etc.) and domain independent ontology like PROTON [11]. However, the level of complexity and the amount of time, it takes to refine the classes and their relationship with external sources of data (e.g. concepts disambiguation, over linking, word sense and terms stemming), leaves web scalability as a non-addressed issue. All such approaches tend to distort the actual users queries [12] thus turning the words in ambiguous queries leading to less relevant and imprecise search results. Another reason for ruling out extending a general purpose light-weight ontology like PROTON [13] to include new classes and properties in a domain ontology along with defining and applying subsumption rules, is the lack of experts and participants (to continuously evaluate new classes) in a scientific field such as Social Sciences. Moreover continuously monitoring the emergence of new concepts and terminologies in a particular domain specific ontology like Social Science Research Methods is in itself unsustainable and prone to ambiguities at every stage.

On the another hand, [14] proposes key-phrase extraction based on semantic blocks which entails pre-selecting blocks of information that have higher coherence in terms of extracting the most meaningful key terms from a web page. Such an approach further complicates the already presumptuous approach of ontology-based entity and concepts extraction by adding another assumption that a more coherent space in a web page or web documents will be preselected before annotating the content inside.

Ontology availability, development and evolution make it a hard choice for developers who are mainly responsible for the implementation of semantic search web applications. We recognize that KIM (Knowledge & Information Management) offers a running platform for ontology-based annotation and retrieval [15] but amalgamating the built in KIM ontology with domain specific ontology followed by gazetteer-based annotations require huge efforts both on the part of developers and ontology designers. KIM server-based search application is still far from being implemented in typical client server architecture, which is what we have experimented with as part of this research.

3 Our Methodology

We describe a framework that incorporates semantic text analysis of content in ReStore repository to add semantic metadata and topical keywords to build a keywords and semantic Knowledge Base (KB). To address the issue of ambiguous terms during annotation, extend the semantic meaning of concepts in web pages and sustain the meaning of terms and concepts in scientific repositories with the passage of time, we have adopted Linked Open Data (LoD) as a tagging data source for various types of documents in our repository. The LoD numbers over 200 datasets which span numerous domains such as media, geography, publications and life sciences etc. incorporating several cross-domain datasets [16]. It is an open source of structured data, which so far has been employed for building Linked Data (LD) browsers, LD search engines and LD domain-specific application such as semantic tagging [17]. A number of web services have been developed recently to extract structured information from text (incorporating LoD) such as Alchemy API[5], DBPedia Spotlight[6], Extractive[7], OpenCalais[8] and Zemanta[9]. We have used Alchemy API due to its holistic approach towards text analysis and broad-based training set (250 times larger than Wikipedia) used to model a domain like ours. The tool uses machine learning and natural language parsing algorithms for analyzing web or text-based content for named entity extraction, sense tagging and relationship identification [18]. Alchemy API was also one of the best in the performance evaluation review of [19] where Alchemy API remained the primary option for NE recognition and overall precision and recall of NEs and types inferences and URI disambiguation.

[5] http://www.alchemyapi.com.
[6] http://dbpedia.org/spotlight.
[7] http://extractive.com.
[8] http://www.opencalais.com.
[9] http://www.zemanta.com.

We start by making the case for information retrieval from the KB and build on this by evaluating search results in terms of TREC[10] 11-points Precision/Recall measures to assess accuracy in search results. We also measure users' tanking to assess the level of users' satisfaction (degree of relevance), which extends the meaning of relevance from binary (relevant or non-relevant) evaluation to users' satisfaction evaluation. We describe the framework as comprising of four elements i.e. (a) Semantic structuring of ReStore repository's content (Schematization); (b) Mass annotation of content addressing all types of content (static/dynamic); (c) Indexing semantic annotation along with actual content (Semantic annotation for metadata augmentation); and (d) Web-based search application using Elasticsearch distributed searching (for precision/relevance evaluation in search results).

3.1 Annotation of Heterogeneous Content with LoD

As it can be seen in Fig. 2 below, we have setup an environment where content from two web sites i.e. www.restore.ac.uk and www.ncrm.ac.uk are extracted by using two distinct methods i.e. crawling static web pages from the websites and extracting dynamic content and non-web page documents from database management systems e.g. MySQL. Figure 2 shows our four framework components i.e. schematization, annotation, indexing, and retrieval in action. Semantic annotation of web documents is performed using one of the best semantic annotators i.e. Alchemy API which analyze each document by using built in NLP (Natural Language processing) and Machine Learning (ML) and other complex linguistic, statistical, and neural network algorithms.

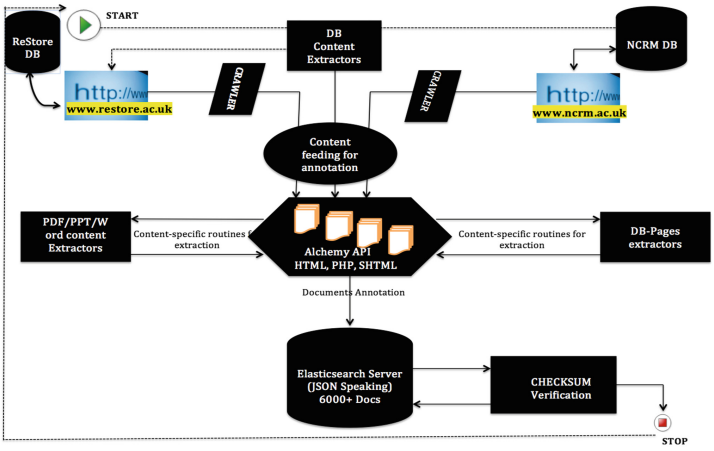

Fig. 2. Web repositories extraction, annotation and indexing process flow

This service crawls billions of pages every month thus expanding its knowledge base through which entities and concepts are identified in web documents and linked to

[10] Text Retrieval Conference http://trec.nist.gov/.

various linked data sources of data. We obtained special license from the Denver-based company allowing us to analyze 30,000 documents per day. Similarly we have added topical keywords, concepts and entities to 3000+ documents and have stored the relevant TF.IDF score along with Alchemy API score against each of the individual items in a single record to enable ourselves to manipulate precision and recall during various experiments and evaluation exercises in later sections. We have indexed data in multiple indexes based on the type of data, size of documents and the possibilities in which the indexed data could be searched and browsed. We have achieved this by designing unique schemas for each index. Finally, we have developed a searching application, which sit on top of the above to facilitate users search for their topic of interest as part of evaluation exercises.

3.2 Scope of Annotation and Indexing

The scale, at which we have setup the annotation, indexing and search results evaluation environment, is extensible (for example beyond ReStore content) and offer greater degree of freedom in terms of utilizing annotation APIs e.g. Concepts, Entities and Keywords. We can for instance use sentiment analysis in order to determine the degree of relevance of concepts, entities and keywords in our documents across different LoD datasets. That in essence offers various search performance levels which could be evaluated at the time of searching to attain the required level of performance. We have for example evaluated the relevant concepts, tagged by our system, by presenting most relevant tags to evaluators next to each search results link as detailed in the evaluation section of this paper. Unlike evaluating Ontology classes of a specific domain, assessing the appropriateness of a sub class in a domain Ontology before annotating content with it, we are interested in annotating topic of interests in a range of topically diverse heterogeneous content contained in web pages, portable PDF documents, presentations and other file formats. To assess the appropriateness of annotation, we present search results to evaluators based on topic (term, concept, keywords) popularity and weight of relevance calculated at the time of annotating and indexing.

4 Implementation

To start off, we have subdivided the annotation process into two distinct categories i.e. 1. Annotation of content based on topic keywords using Alchemy Keywords API and 2. Semantic annotation of content using Alachemy Concepts, Entities APIs. The content includes web pages, PDF, CSV, Word, Powerpoint presentations and other software code script produced as part of various experiments by the researchers during the course of research projects.

4.1 The Annotation Process

Figure 3 shows the entire annotation process along with web-based search interface for evaluation purposes. The process flow starts from the diamond-shaped box which is

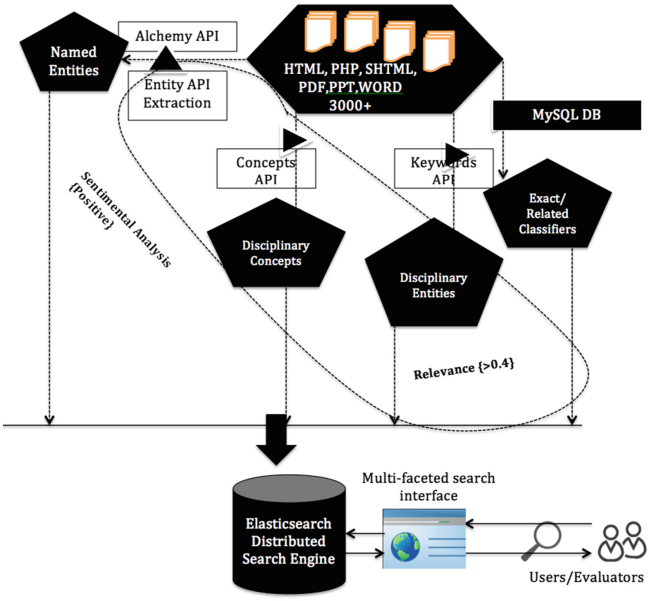

Fig. 3. Automatic annotation of content extracting keywords, entities and concepts using Alchemy APIs (incorporating LoD datasets i.e. DBPedia, YAGO [20], OpenCyc (OpenCyc contains hundreds of thousands of general knowledge terms organized in a carefully designed ontology. Available at http://opencyc.org/doc/opencycapi) and Freebase [21]) keywords entity and concepts extraction

our benchmark document corpus hosted by two LIVE websites i.e. NCRM and ReStore.

By using Alchemy API service, we take the entire corpus of structured documents as a Knowledge Base which conform to the data retrieval model elaborated by [1]. With semantic annotation two important tasks of the semantic web can be achieved i.e. (1) extracting and hyperlinking named entities in documents and (2) finding relevant documents in accordance with entities [22]. All this brings structure to web content in order to enhance the meanings of text in web pages, which improve content searching based on mutual relationships between different sections of documents, keywords and entities. Based on this interpretation, we assume that (a) we have an entity or context extraction platform which would be applied on a number of documents in order to pinpoint keywords, entities and concepts to a more meaningful contemporary source of data e.g. DBPedia, Freebase and Yago; and (b) build a Knowledge Base of data which comprise actual documents and semantic metadata along with inter-documents relationships.

4.2 Elasticsearch Search Engine as a Knowledge Management Platform

We have used Elasticsearch server mounted on the current ReStore web server with an intention to turn it into a full-fledged dedicated semantic and full text search server used

with the front end (PHP, JavaScript, Elastica Library) to display search results to users for evaluation purposes. We have addressed a particular issue which many semantic search system suffers from i.e. usability limitations where users are expected to use formal query language to express their requirements and lack of optimal semantic annotation of content in web documents emanating from using small set of pre-defined domain ontologies and datasets [1]. Using DBPedia spotlight discussed in [23], for instance assumes that users should be able to opt for preferred or alternative labels while searching for things in the DBPedia spotlight web application. We understand, that such assumptions compromises the soundness of semantic data-based web application as the majority of users still prefer to use free text keywords based search without pre-specifying advanced search options [9].

Elasticsearch analyzers first analyze all the content belonging to each document via JSON-formatted URLs and relevant scores are stored against keywords, entities and concepts (extracted by Alchemy API using three different API services i.e. Keywords, Entity and Concepts). Each document D_j represents a vector space model in the following manner:

$$D_j = (t_k, t_e, t_c \ldots, t_{kec})$$

Where t_k, t_e, t_c, t_t are the keywords (k), entities (e) and concepts (c) terms. With this representation, each document is a vector having the above elements for influencing ranking of search results. The scoring algorithms are based upon statistical and NLP techniques employed by Elasticsearch distributed search server. It is however to be mentioned here that we have indexed individual Alchemy APIs-based score as well in each index in order to run sub queries based on users' browsing preferences but browsing based evaluation is beyond the remit of this paper.

5 Evaluation

We have analyzed the performance in terms of search results relevance, precision and recall in two different categories i.e. (1) searching on the basis of keywords and actual content (2) searching on the basis of topic keywords, semantic concepts and entities extracted by Alchemy Keywords, Concepts and Entities APIs. In our benchmark document collection, we have annotated more than 3000 documents and have built a semantic store ready to be exploited by our web-based search application. The whole process is carried out in a client-server architecture which includes ReStore web server, ReStore and NCRM Database server containing 5 different MySQL Databases populating almost 6000 web pages in NCRM website and 3000 in ReStore.

Similarly in our queries benchmark collection, we have captured a set of free-text queries from Google Analytics which were submitted by online users of the ReStore website to find the desired information using the current web search. The total number of queries randomly selected was 34. The criteria for selecting these queries included the number of clicks they generated and brought users to the ReStore repository site, recency of use and meaningfulness of terms in the query. However, to reduce bias in all 34 queries, we selected queries having one single term, multiple terms and mixture of

keywords and concepts. Here is a sample of user queries in the benchmark query collection. {*cohort sequential design, design effects in statistics, paradata in survey research, randomized control trials, evaluating interaction effects, ethnic group, Forecasting, Stages of a systematic review, sample enumeration, what is media analysis.}*

With regard to the experts' judgment evaluation, our evaluators included Librarians, Social Science academics, Social Science research fellows and PhD students coming from various disciplines e.g. Social Sciences, Education, Geography & Environment and Statistics. Feedback was collected from 15 expert evaluators over a period of 3 weeks. The evaluation exercises were designed in such a way that they had a freedom in reviewing 70 web documents at their own pace as long as they were logged in. We asked the evaluators to carry out search exercises by using the pre-selected set of queries. Their assessment included whether (a) a search result is relevant or not after viewing the content of the results by clicking on the link; and (b) ranking the result in terms of the number of stars corresponding to each results; and (c) authenticating potential concepts/entities from the list having association with each result. The first 10 results retrieved were assumed to be relevant against each query. Each expert user was given a set of 7 queries before logging on to the search results evaluation page and his/her activity was recorded in the database along with their ranking of each search results. It is to be mentioned here that some queries were evaluated by only two evaluators in which case we averaged the ranking before using it in our evaluation analysis. On the basis of their assessment, we have computed MAP (Mean Average Precision) and have drawn TREC's (Text Retrieval Conference)[12] points recall precision curves in the next section to show that enhanced semantic metadata attachment to the actual content clearly improves precision in search results with maximum recall. We have also computed ranking performed by users (MAR-Mean Average Ranking), which reflects the system's ranking in terms of accuracy and results relevance.

5.1 Search Results Evaluation

We expected each participant to evaluate 70 search results in total i.e. 10 against each query (7 queries per participants). All together 15 expert evaluators evaluated the system and 886 web documents were evaluated. These participants also added 2555 semantic concepts and entity tags with these 886 web documents. Our system presents a list of 10 results to users with a summary for each highlighting matched words in the query, which helps users make a quick sense of the result before clicking on it. We assume that typically every web users would want every result on the first page to be relevant (high precision) but have little interest in knowing let alone looking at every document that is relevant. We have used precision/recall measures to determine the system performance. Recall is the ratio of relevant results returned divided by all relevant results and precision is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant results retrieved.

While calculating query-level precision, recall and Average Precision (AP) we assumed all top 10 documents retrieved against each individual query to be relevant. We calculated Average Precision (AP) on query level using TREC's 11-points recall

and compared it with that of keywords-based AP to ascertain which system performed better in terms of precise search results without compromising too much recall. To properly quantify the level of relevance in both scenarios, we have used the combined measures that assess the precision/recall tradeoffs which is given by:

$MAP = (\sum_{i=1}^{Q} AP_i)/Q$ where Q = number of queries in a batch. We have also used MAR (Mean Average Ranking) to ascertain the degree of relevance in terms of users' happiness based on users' ranking.

5.2 Semantic Entities and Concepts Tagging

Alongside actual results evaluation and ranking of each result, participants also tagged relevant concepts and entities, presented to them by the system next to each result. The tagging has been of help in understanding participants' decision element for ranking a particular result. For example when "forecasting" is searched, one of the concepts that was suggested to users for tagging in a few results was "prediction", "decision theory", "Bayesian inference" and "statistical inference". Those relevant concepts had already been identified by the semantic annotator but participants' tagging enabled us to re-validate the system's accuracy, which is reflected in assessing the degree of user happiness or MAR in Fig. 6. In contrast, when forecasting was searched in the existing online search facility of ReStore, most of the results in top 10 results, were retrieved because of the mention of the word forecasting. Likewise, when forecasting was searched by multiple users as part of our evaluation, they highly ranked a result which was no 3 in the top 10 list and it had no mention of forecasting but the content were about a research tool used to predict housing, income and education situations of participants taking part in a case study. Google's top 10 results included those defining forecasting, Meteorological office forecasting and baseball game forecasting.

The better performance in semantic search results evaluation in Figs. 4 and 5 is attributed to the semantic index scoring criteria. For example, one of the queries *mixture model"* doesn't exist in the document vector space under the "keywords" list but it has a high score under the "concepts" list of that document. Similarly another query "*Randomized control trials*" exists in the document space under "keywords" list but with a low score and high score under the "concepts" list. Thus when searched, the document having such concept (not necessarily under keywords) with high score was presented to user as highly relevant. Likewise, when "*reasoning*" was searched, a document containing "*critical thinking*" concept came first in top 10 search results which was tagged as relevant by the evaluators.

5.3 The Experiments

After entering individual queries in the search box, a participant was expected to classify a result as either relevant or irrelevant i.e. 1 for relevant and 0 for irrelevant web documents. The participant also star-rated the result and tagged relevant concepts and entities retrieved along with individual results, which can be used for measuring average ranking across the set of queries.

For instance we have to see how many relevant pages r = {r^1,r^2,…r^n} could be retrieved in top 10 pages which were retrieved against each query from keywords index Q(k) = {k^1, k^2, k^3…k7} and semantic index Q(s) = {s^1, s2, s3…s^7}. In other words, how best could our system interpret keywords in users' queries, turn them into topical keywords, concepts and entities and retrieve those web pages, annotated by the annotator during the annotation and scored at the time of indexing. Precision at *k* documents has been our assumption throughout the experimentation process, which implies that the best set of search results appears on the first page of results set and the total number of best results is 10. Similarly recall at *k* documents is based on our assumption that the relevant documents at the time of submitting each query will remain 10 documents. This approach has been adopted based on the nature of web searching in ReStore repository which host archived content and most users are by and large interested in the first 10 results to maximize their satisfaction in terms of finding relevant results.

5.4 Fixed vs. Interpolated Precision Measurement

We assume during our evaluation that a user will examine a fixed number of retrieved results and we will calculate precision at that rank and interpolated rank. Hence our fixed average precision is given by: $P_{(n)} = \sum_{n=1}^{N} \frac{r(n)}{n}$ where r(n) is the number of relevant items (at cut off k relevant document) retrieved in the top n which in our case is 10 documents (N) at each level of individual information needs in the form of user queries. However, if *(k + 1)th* retrieved is not relevant, precision will drop but recall will remain the same. Similarly if *(k + 1)th* document is relevant both precision/recall increase. Therefore we had to extend these measures by using ranked retrieval results, which is a standard with search engines. Interpolated precision is therefore given by:

$$\mathrm{P}_{11-\mathrm{pt}} = \frac{1}{11}\sum_{j=0}^{10}\frac{1}{N}\sum_{I=1}^{N} P_i(r_j)$$

where $P(r_j)$ is the precision at our recall points but $P(r_j)$ doesn't coincide with measurable data point *r* if number of relevant documents per query is not divisible by 10 in which case, the interpolated precision is given by:

$\mathrm{P}_{interpolated}(r) = \max\{P_i : \mathrm{r}_i \geq r$ *where (Pi, ri)* are raw values obtained against different queries or information needs. So the new average interpolated precision is given by:

$$\mathrm{P}_{11-\mathrm{pt}-\mathrm{interpolated}} = \frac{1}{11}\sum_{r\in\{0,0.1,\ldots,1.0\}} p_{interp}(r)$$

In other words, interpolated precision shows maximum of future precision values for current recall points. The normal precision/recall curve reacts to such variations differently as we have shown in Fig. 4. An increase in both precision and recall means, the users is willing to look at more results. All this tells us about expected precision/recall values for another set of results (k + 1,2,3…n). Since $\mathrm{P}_{(n)}$ ignores the

rank position of relevant documents retrieved above cut off (i.e. 10 + 1), we have calculate interpolated precision at each query level to assess the system performance at n + 1 documents which is beyond the existing cut off point.

Figure 4 shows the TREC-11 points ranked retrieval precision/recall curve which is representative of the two systems we have assessed i.e. topical keywords or full-text search vs. semantic index-based searching.

Figure 4 shows TREC 11 points Interpolated Precision and Recall curve showing system performance i.e. keywords (full-text) searching vs. keywords & semantic index-based searching. It also shows averages system's performance over the entire queries batch.

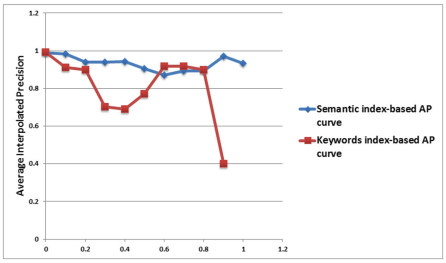

Fig. 4. TREC 11 points Interpolated P/R

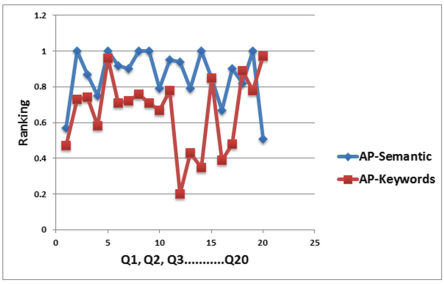

Fig. 5. Non-interpolated AP curve

We can clearly see that the behavior of un-interpolated keyword index-based precision-recall curve is quite fluctuating while that of semantic index remains firm in interpolated Fig. 4, vacillating between 100 and 80 % for the first top 10 results for the entire queries batch. Of course we have given some ground to the fact that only 20 out of 34 queries were attempted by two different types of evaluators i.e. a batch of 7 queries was attempted by two evaluators one based on full-text search index and another semantic index. We also averaged multiple queries in order to get the overall performance of keywords based and semantic index-based search system. We also calculated non-interpolated average precision at each query level which is given by: $P(r) = \sum_{i=1}^{N_q} \frac{P_i(r)}{N_q}$ where $P(r)$ is the average precision at Recall level r and N_q is the number of queries. $P_i(r)$ is the precision at Recall level r for the i-th query. Figure 5 clearly shows better performance in terms of un-interpolated precision and recall when queries were searched against semantic index. Performance in both situations suggests that semantic index based curve performs better. Figure 5 indicates overall better performance with the exception of few queries where keywords index-based curve performs better. But this has been offset by the interpolated precision-recall curve in Fig. 4 where semantic-index-based performance remains consistent.

We have also calculated Mean Average Precision (MAP), which has become commonplace in recent years providing a single-figure measure of quality across recall levels. Using MAP, fixed recall levels are not chosen and there is no interpolation. $\text{MAP} = (\sum_{i=1}^{Q} AP_i)/Q$ where Q = number of queries in a batch. MAP ensures that equal weightage is given to all queries i.e. those containing rare and common terms

with different recalls. Our MAP for keywords and semantic search results are 66 % and 84 % respectively.

5.5 Mean Average Ranking

Precision and Recall curve don't allow as such for the degree of relevancy when it comes to retrieving precise and relevant documents. This is partly because this measures is based on binary classification and individuals' perception. What is relevant to one person may not be relevant to another. To address this issue, we have averaged ranking of all relevant documents against each of all queries in our experimental batch and have shown the average ranking in the following graph to prove our point. We assume that along with measuring the degree of relevance in search results, it is also equally important to measure the utility or satisfaction level achieved by users after exploring through the actual content. Mean Average Ranking (MAR) therefore represents our ranking model, which have applied in the following diagram.

$MAR = (\sum_{j=1}^{Q} AR_j)/Q$ where Q = number of queries in a batch.

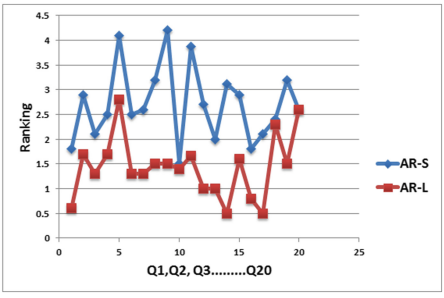

Fig. 6. Average ranking curve shows averages ranking over a medium set of queries. We have computed average ranking at each information need or query level and have concluded that MAR in semantic index-based searching performs better than the full-text index

6 Conclusion and Future Work

The aim of this research has been to explore new avenues for semantic index-based searching and assessment of users' happiness at the time of searching for relevant results. By correct identification of LoD-based topical keywords, concepts and entities, users could continue finding relevant information using online search regardless of the time factor. In other words by regularly adding semantic metadata using LoD, content will continue to sustain its value through enhanced semantic metadata annotation and indexing using online search applications regardless of the time and discipline elements. To advance this research, we will further investigate users' contribution or crowd-annotation (using controlled vocabulary and free text tags) at the time of exploring content web repositories and its impact on search results. Using built-in annotation tools in every web page of the ReStore and NCRM repositories

(already deployed in benchmark documents), we will conduct focus group exercises to (a) crowd-annotate various content in the above online repositories to address concepts obsolescence; and (b) sustain the quality of search results regardless of time factor in the designated repository of social science research data.

References

1. Fernandez, M., et al.: Semantically enhanced information retrieval: an ontology-based approach. Web Semantics: Science, Services and Agents on the World Wide Web **9**(4), 434–452 (2011)
2. Wu, P.H., Heok, A.K., Tamsir, I.P.: Annotating the web archives – an exploration of web archives cataloging and semantic web. In: Sugimoto, S., Hunter, J., Rauber, A., Morishima, A. (eds.) ICADL 2006. LNCS, vol. 4312, pp. 12–21. Springer, Heidelberg (2006)
3. Khan, A., Martin, D., Tiropanis, T.: Using semantic indexing to improve searching performance in web archives. In: International Journal on Advances in Internet Technology, Seville, Spain, pp. 1–4 (2012)
4. Riggs, F.W., Interconcept report: a new paradigm for solving the terminology problems of the social sciences, UNESCO, vol. 44 (1981)
5. Snow, R., et al.: Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (2008). Association for Computational Linguistics
6. Royo, J.A., et al.: Searching the web: from keywords to semantic queries. In: Third International Conference on Information Technology and Applications, ICITA 2005 (2005)
7. Zervanou, K., et al.: Enrichment and structuring of archival description metadata. In: ACL HLT 2011, p. 44 (2011)
8. Benjamins, R., et al.: The six challenges of the semantic web (2002)
9. Yang, C., Yang, K.-C., Yuan, H.-C.: Improving the search process through ontology-based adaptive semantic search. The Electronic Library **25**(2), 234–248 (2007)
10. Georgiev, G., et al.: Adaptive semantic publishing. In: WaSABi@ ISWC (2013)
11. Damova, M., et al.: Mapping the central LOD ontologies to PROTON upper-level ontology. In: Proceedings of the Fifth International Workshop on Ontology Matching (2010)
12. Shabanzadeh, M., Nematbakhsh, M.A., Nematbakhsh, N.: A semantic based query expansion to search. In: 2010 International Conference on Intelligent Control and Information Processing (ICICIP) (2010)
13. Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D., Goranov, M.: KIM – semantic annotation platform. In: Fensel, D., Sycara, K., Mylopoulos, J. (eds.) ISWC 2003. LNCS, vol. 2870, pp. 834–849. Springer, Heidelberg (2003)
14. De Virgilio, R.: RDFa based annotation of web pages through keyphrases extraction. In: Meersman, R., et al. (eds.) OTM 2011, Part II. LNCS, vol. 7045, pp. 644–661. Springer, Heidelberg (2011)
15. Bai, R., Wang, X.: A semantic information retrieval system based on KIM. In: 2010 International Conference on E-Health Networking, Digital Ecosystems and Technologies (EDT) (2010)
16. Rusu, D., Fortuna, B., Mladenic, D.: Automatically annotating text with linked open data. In: LDOW (2011)
17. Bizer, C., Heath, T., Berners-Lee, T.: Linked data-the story so far. In: Semantic Services, Interoperability and Web Applications: Emerging Concepts, pp. 205–227 (2009)

18. Gangemi, A.: A comparison of knowledge extraction tools for the semantic web. In: Cimiano, P., Corcho, O., Presutti, V., Hollink, L., Rudolph, S. (eds.) ESWC 2013. LNCS, vol. 7882, pp. 351–366. Springer, Heidelberg (2013)
19. Rizzo, G., et al.: NERD meets NIF: lifting NLP extraction results to the linked data cloud. In: LDOW, vol. 937 (2012)
20. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: Proceedings of the 16th International Conference on World Wide Web, pp. 697–706. ACM, Banff, Alberta, Canada (2007)
21. Bollacker, K., et al.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of data, pp. 1247–1250. ACM, Vancouver, Canada (2008)
22. Kiryakov, A., et al.: Semantic annotation, indexing, and retrieval. J. Web Semant. **2**(1), 49–79 (2004). Elsevier's
23. Mendes, P.N., et al.: DBpedia spotlight: shedding light on the web of documents. In: Proceedings of the 7th International Conference on Semantic Systems, ACM (2011)