Exploiting semantic annotation of content with Linked Data to improve searching performance in web repositories

Arshad Khan[Ω], Thanassis Tiropanis[ϕ] & David Martin[*]

School of Electronics & Computer Sciences, University of Southampton, UK

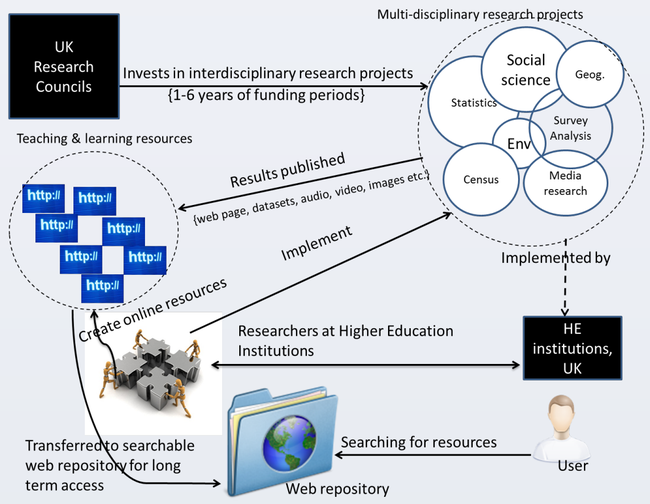
Introduction

Searching for relevant information in web repositories of multi-disciplinary 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 producing the desired search results due to 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 lack of trust in such repositories of valuable knowledge.

This research explores the appropriateness of alternative searching based on Linked Open Data (LoD)-based semantic annotation and indexing in online repositories such as the ReStore repository (www.restore.ac.uk) containing content from multiple Social Science research methods projects . We explore websites content annotations using LoD to generate contemporary semantic annotations. We investigate whether we can improve accuracy and relevance in search results affected by concepts and terms obsolescence in repositories of scientific content.

Overview

Multidisciplinary research teams publish their findings in various forms including blogs, personal web pages and project websites.



- Web resources publication and archiving in web repositories has never been easier but finding relevant content along with the growth of repositories is a challenge.
- A lack of metadata or annotation at the time of storage/archiving adversely affects search results.
- Unambiguous and contemporary metadata annotations are needed to cope with the challenges associated with finding relevant information.

The Problem

- Current searching techniques in web repositories are predominantly based on keywords instances with very little or no attention paid to semantic meaning, types of content, context and relationship of keywords/phrases in web pages.
- Change in scientific terms and concepts is inevitable due to rapid growth of knowledge and research. This change occurs over the passage of time, influenced by factors like cultural, social, technological, scientific and socio-economic changes.
- Ontology-based semantic expressions and the structuring of content by matching terms with the resulting ontology classes has been around for over a decade.
- However, frequent intervention is needed by ontology developers and subject experts aimed at disambiguation and syntactic/grammatical correctness.

Our research

We have focused on 3 main areas:

1. Whether obsolescence in terms and concepts could be addressed by semantic annotation, thus enriching the keywords index with meaningful metadata?
2. Whether a shift from domain-specific ontology-based semantic annotation to distributed and wide data spaces like Linked Open Data (LoD)-based semantic annotation could address the issue of entity, concepts and relation disambiguation.
3. How practical would it be to design, build and deploy a scalable web-based annotation, indexing and retrieval system aimed at continual semantic annotation of heterogeneous content, indexing and searching based on changing trends, specifically in a Social Sciences context.

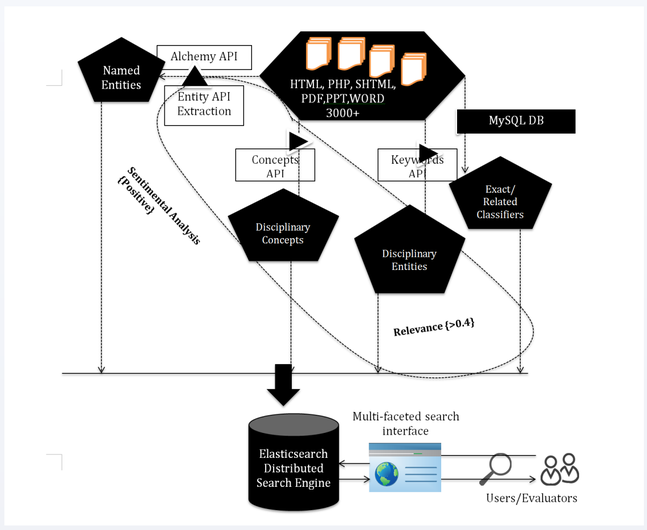
Implementation

We have built a fully-fledged system for evaluating searching performance in various situations using Elasticsearch, PHP, MySQL and JavaScript. There are 3 main components of the evaluation system:

1. Annotation of content (Web pages, PDF, PPT, TEXT etc.) to extract Named Entities, Concepts and Topical keywords using LoD datasets e.g. DBPedia, YAGO and OpenCyc.
2. Schematization and indexing of content and annotations in Elasticsearch distributed search platform
3. Web-based searching as part of evaluation

We have used Alchemy API which analyzes each document by using built in NLP and Machine Learning and other complex linguistic, statistical and neural network algorithms and extract named entities automatically.

A system diagram, showing the annotation, indexing and searching of various types of content in ReStore repository, further elaborates various component of our system.

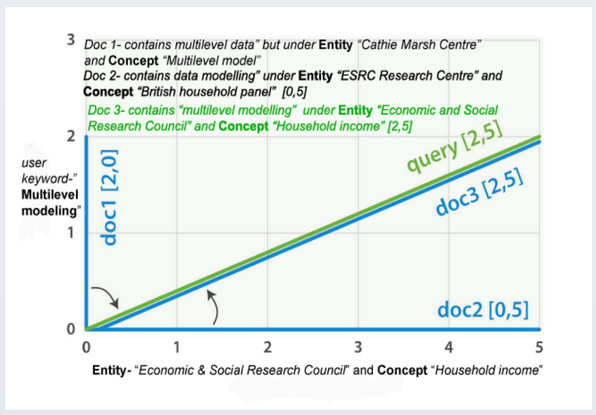


Overview of LoD-based semantic annotation of various types of content and search results retrieval

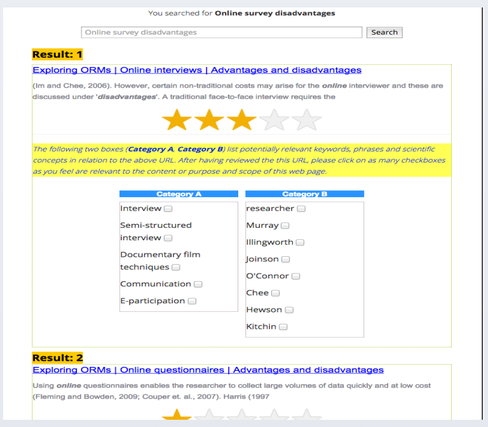
Experimentation

Participants in experimentation were asked to classify web documents resulting from a search as either relevant (1) or irrelevant (0). The participants were also asked to star-rate the result and tag relevant concepts and entities retrieved along with the individual results. This was used for measuring average ranking across the set of queries. Further details include:

- Benchmark query collection contained 34 queries pre-selected from Google analytics
- 15 expert evaluators were recruited for the evaluation exercises
- Each evaluator (Post-doc researchers, lecturers, PhD students) reviewed 70 documents retrieved against 7 benchmark queries. 886 web documents in total were evaluated and tagged with extra concepts/entities.
- A web-based user interface was used for search results retrieval and evaluation which included star-rate classification and tagging of search results (Right)
- Information need (query) vs. desired results model is given below (Left)



Conceptual framework for information needs & results

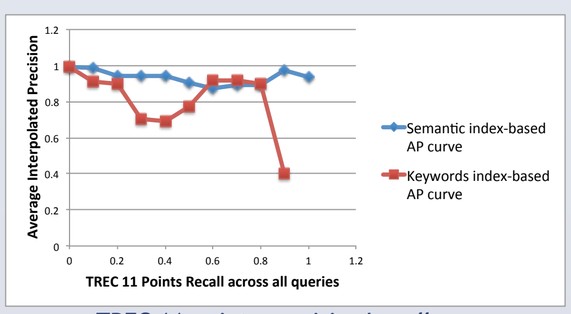


Results evaluation interface

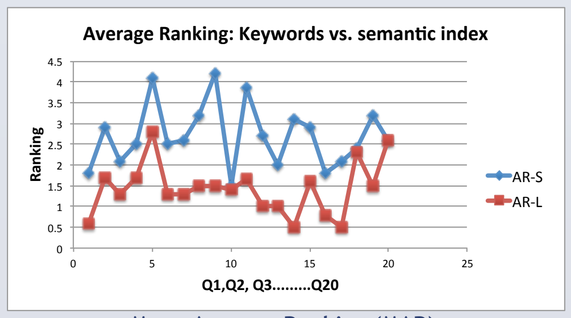
Evaluation & Future work

We have computed MAP (Mean Average Precision), TREC-11 points Precision/Recall and Mean Average Ranking (MAR) to measure system performance in terms of optimised search results. The first 10 results were used as evaluation metrics for all queries in each evaluator case. Search results were produced based on TF-IDF score computed for two distinct groups of participants i.e. those who searched across 3000+ ReStore repository documents using:

1. Keywords + content based searching
2. Searching on the basis of topic keywords, semantic concepts and entities



TREC 11-points precision/recall curve



Mean Average Ranking (MAR)

- MAP (full-text index)=66%, MAP (Semantic index)=84%
- TREC-11 points Interpolated Precision shows system performance in terms of keywords-based searching and semantic annotation-based searching. The precision-recall curve remains stable between 80% and 100% precision across the entire batch of queries.
- The figure (left) also shows maximum future precision values for current recall points.
- The figure helps us predict that user is willing to look at more results beyond the first 10 which is our benchmark for the results evaluation.
- MAR figure (right) shows degree of relevancy for the retrieved &relevant documents
- Our findings clearly show that enriching full-text index with contemporary LoD-based semantic annotations improve precision, relevance in search results and elevate user satisfaction. We aim to extend this research by incorporating Crowd-annotations and Social Research classification vocabulary for tagging content to improve search results.

Acknowledgement

The authors acknowledge the support of ESRC† award no. RES-576-25-0023 and ES/L008351/1. We also acknowledge the support extended to us by the Alchemy API team for their free license to annotate website contents (30, 000 pages per day).

Ω Arshad Khan is a PhD candidate at ECS and works for the National Centre for Research Methods (NCRM), University of Southampton
φ Thanassis Tiropanis is Arshad Khan's PhD supervisor based in ECS, University of Southampton
* David Martin is based in Geography and Environment, University of Southampton

Author's email: aak1v11@ecs.soton.ac.uk

† Economic & Social Research Council (ESRC)