Semantic Enrichment for Local Search Engine using Linked Open Data

Mazen AlObaidi
Oakland University
malobaid@oakland.edu

Khalid Mahmood
Oakland University
2200 N. Squirrel Rd
Rochester, MI 48309
+1 248-370-3542
mahmood@oakland.edu

Susan Sabra
Oakland University
sabra@oakland.edu

ABSTRACT

Local search engines are a vital part of presence of local businesses on the Internet. Local search engines improvement is an important element to ensure that local businesses can be found by millions of people whom are using web to find services. However, web presence can be disguised or not properly documented. Our approach will improve the effectiveness of local search, and increase the ranking of local businesses. We introduce an approach for enhancing local search engines efficiency in returning more accurate results. Our approach consists of semantically enriching the results of a query using Linked Open Data (LOD) web content. Our preliminary evaluation demonstrated evidence that our approach has a better search with more accurate results.

Keywords

Semantic enrichment, linked open data, knowledge discovery, triple extraction, RDF.

1. INTRODUCTION

Open Web Data constitute a very rich environment for searching and retrieving information. Local search engines are a very common tool for finding location-based services available on the open web today. However, a possible problem with these services can occur when users' requests into a local search engine receive an outcome that might not necessarily be the best or the optimal one to make a decision. For example, when a user looks up for a specific type of labor such as a local plumber or a local painter in a local search engine, the results of that search might not be as accurate as the full listing because some individual laborers might not be registered for online business. Another example can be a small business that provides a service but it is not listed on the top sub list to be advertised as such. We propose a feasible solution for this problem of local search engine using semantic enrichment based on Linked Open Data (SE-LOD). We propose a new approach to semantically enrich the result of a local search engine through matching LOD resources that would improve the indexing of the results. In addition, the new list of results will have a better indexing than a typical local search engine. The new indexing will be enhanced due to the additional information semantically related to the query, found on the open web. Our main contribution takes two major points: firstly, we propose a new way for using LOD in semantic enrichment of query in local search engine, and secondly, a subsequence of the first contribution is the enhanced indexing in the local search engine for providing more satisfactory and better results.

2. RELATED WORKS

The proliferation of the use of the Web of Data and the increasing amount of Linked Open Data (LOD) constitute a rich environment that stimulates the development of new applications and the wider consumption of published data [6]. The LOD cloud makes it possible to semantically enrich unstructured user-generated content with structured information presented in the LOD resources. In [10] they studied the extent to which the Linked Open Data cloud can help to semantically enrich volunteered geographic information in order to better answer queries in the context of crisis and disaster relief operations. Wetz et al. in [11] used LOD to semantically enrich thesaurus concepts by comparing the similarity algorithms used in finding matching of LOD concepts to the local thesaurus. Bontcheva et al. in [3, 4] used LOD semantic enrichment for environmental science literature for metadata and full text articles in order to enhance the search process. De Faria Cordeiro et al. in [5] proposed a semantically enriching approach for governmental data approach to support the exposure, sharing and association of data resources in the form of Linked Open Data, offering a user-friendly environment to stimulate the publication of data and their association to other existing data. In [2] they used semantic annotations to enrich the document metadata and provide new types of visualizations in an information retrieval context for identifying bibliographic references in texts. In [6] they describe an approach for generating and capturing Linked Open Provenance (LOP) to support data quality and trustworthiness assessment.

3. PROPOSED APPROACH

3.1 SE-LOD Framework

We propose an approach that enriches the web content of a local search engine. The primary goal of our approach is to make local search indexing more effective and efficient, which would in turn improve the rankings of local businesses. Our framework (see Figure 1) contains four components: the "Natural Language Processing Module" (NLP), the "Name Entity Recognition," the "Entity URI Look-Up," and the "Triples Extraction". Moreover, our approach consists of two steps: "entity segmentation" and "knowledge extraction." The two steps are critical to realize the primary goal of our approach.

WWW 2016 Companion, April 11-15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2890481

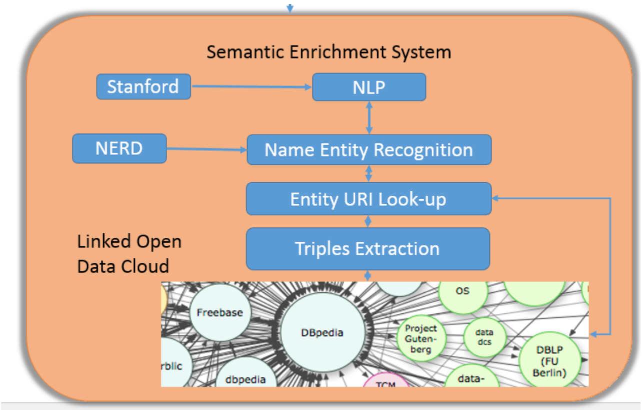


Figure 1: SE-LOD Framework

3.2 Segmentation

Segmentation is both the process of tokenization and the process of identifying name entities. The processes begin with empty, raw text that is highlighted as sentences. These sentences are mere raw material. They have potential, but there is little semantic use to be found there. The process by which tokenization and the recognition of name entities occurs follows according to two modules:

3.2.1 Natural Language Processing

The Natural Language Processing Module is a data processing center that plays a central role in SE-LOD architecture. This module uses the Stanford CoreNLP functionality in order to conduct deep sentence analysis that produces a collection of lists of a structure tree. The next step is to take these lists and stem them so as to actualize the process of reducing inflected words to their root [7]. In addition, the lists are passed on to the name entity recognition module for disambiguation.

3.2.2 Named Entity Recognition and Disambiguation Module

The Name Entity Recognition and Disambiguation Module is given the task of locating and labeling sequences of words in text and categorizing these sequences into a predefined classification. In addition, the module uses specifically Named Entity Recognition and Disambiguation software (NERD) [9]. For example, a name entity such as "Microsoft" will tag and categorize the entity as an organization.

3.3 Knowledge Extraction

The key task of knowledge extraction is to extract knowledge from structured and unstructured data by using Linked Open Data. It also expresses the extracted knowledge in a way that both presents the semantic implicit in that knowledge and facilitates the function of local search engines. Knowledge extraction operates according to two modules:

3.3.1 Entity URI Look-up Module

The literal of each entity that is generated by the segmentation process is also processed by this module. The module identifies related entities from the Linked Open Data program, using properties such as "rdfs:label," "skos:prefLabel," and the "skos:altLabel" properties. The first property is a lexical label for a resource [8], the second is a preferred lexical label for a resource, and the third is an alternative lexical label for a resource [1]. To illustrate, assume that we have an entity called "Microsoft." The module at hand uses the entity label to construct a SPARQL query to extract all the triples that contain the string "Microsoft" as a part of its object. In turn, the module iterates through generated triples, extracts the URI subject, and builds the entity URI list that needs to be used in the next module under which knowledge extracts operate.

3.3.2 Triple Extraction Module

The main task of the module is to accept the entity URI list mentioned above from the Entity URI Look-Up Module, iterates that through the entity URI list, and construct the SPARQL query that uses each URI in the list as the subject, as well as the rdfs:comment that represents a description of the resource as a predicate [8]. Moreover, the module executes the SPARQL query against the Linked Open Data endpoint. Next, the module provides all the comments retrieved from the local search engine so as to consider them in the indexing process (see Algorithm 1). However, the main challenge in this process is improving the Precision. Precision is the ratio of the number of true positive comments retrieved over the total number of comments retrieved while Recall is the ratio of the number true positive comments retrieved over the total number of true positive comments. Knowledge extraction on LOD leads to a high Recall and lower Precision [12].

Algorithm 1

```
Triple_Extraction ( List URIList)
{
  FOR EACH  URI in URIList
    {
      SPARQLString ← BuildSPARQL(URI)
      Rcomment      ← Execute.SPQARQL(SPARQLString)
      RcommentList.add(Rcomment)
    }
RETURN RcommentList
}
```

Therefore, we developed a unique algorithm that improves the precision by filtering the true positive comments out of the retrieved comments by the following process:

1. uses rdfs:type property indicating that a resource is a member of a class
2. iterates through the entity URI list that is mentioned above
3. constructs SPARQL query with subject as entity URI and predicate as rdfs:type
4. executes the SPARQL query against Linked Open Data endpoint
5. builds set of concepts called S given by S={c_1,c_1,….c_n}
6. iterates through the resource comment list that is generated by Triple Extraction call
7. extracts all entities from the raw text “comment”
8. repeat steps 2 to 4 using comment entities that are retrieved in step 6
9. builds set of concepts called SC from the result of step 7 given by SC = {c_1,c_1,….c_m}
10. using overlapping set equations to find the positive comment (PC) as follows

$$PC(S, SC) = (S \cap SC) \geq \text{weight Threshold}$$

4. Evaluation of Semantic Enrichment

One of the essential early steps in the proof of concept process is creating a prototype. In turn, we designed an algorithm and executed an experimental framework using Stanford Natural Language Processing (NLP) APIs [7] which is a program that works out the grammatical structure of sentences by using the parser mash-up. Named Entity Recognition and Disambiguation (NERD) is a framework that integrates ten popular named entity extractors available on the web with white page local search engine, and Linked Open Data, expressing a corner stone of what Semantic Web is all about: large scale integration of, and reasoning on, data on the web.

4.1 Prototype Validation

To simulate a real world example, we used white page local search to find variety of services in local area of Ypsilanti, Michigan. In addition, we manually identified a set of local businesses of the same services. In the next step, we identified the local businesses that are manually found and not part of the local search result. We then passed them into our framework system as input. Then, the generated result from our approach for each input was manually assessed and added to the local businesses that were identified by the white page local search if they matched the service. The primary task of the evaluation is to confirm that our approach is improving the list of local businesses that is returned from local search engine. As an example, we looked up residential local painters in Ypsilanti Michigan using white page local search and Table 1 shows the generated result. Clearly, we noticed that one of the local business called "Home Depot" is not part of the result.

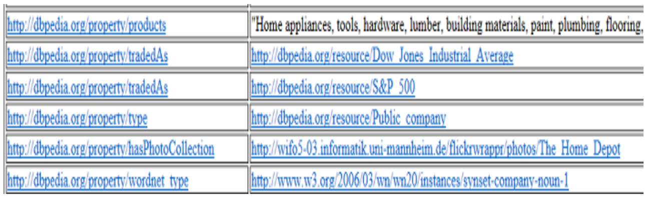
Table 1: Sample regular generated result

Name	Address
Ace Painting	1414 Beard St Huron
Performance Painting	160 Edith Street
All Pro Painting	840 East Maple Road
Dave Kenney's Painting	43373 Ashbury
A G Maintenance, In	325 King Road
Always Painting	Service Ypsilanti Area

Therefore, we have input "Home Depot" business name to our system and assessed the result. Figure 2 shows the result and pointed that Home Depot has painting services.

4.2 Results

Based on our small scale evaluations, clearly we can see there is improvement about average 10 percentage increase in the current local search result. This increase comes from mapping the entities of local search input to semantic linked open data. In addition, scalable local search engine access to linked open data**.**



http://dbpedia.org/property/products	"Home appliances, tools, hardware, lumber, building materials, paint, plumbing, flooring,
http://dbpedia.org/property/tradedAs	http://dbpedia.org/resource/Dow_Jones_Industrial_Average
http://dbpedia.org/property/tradedAs	http://dbpedia.org/resource/S&P_500
http://dbpedia.org/property/type	http://dbpedia.org/resource/Public_company
http://dbpedia.org/property/hasPhotoCollection	http://wifo5-03.informatik.uni-mannheim.de/flickrwrappr/photos/The_Home_Depot
http://dbpedia.org/property/wordnet_type	http://www.w3.org/2006/03/wn/wn20/instances/synset-company-noun-1

Figure 2: Improved result generated after SE-LOD

We conducted an experiment on a variety of services to be queried in a local search engine without any improvement and using SE-LOD. The comparison of results shows an evident improvement change to a 40%. Figure 3 shows in detail the results of our experiment.

5. CONCLUSION

Local search engines are a vital part of presence of local businesses on the Internet. Local search engines improvement is an important element to ensure local businesses can be found by millions of people whom are using web to find services. Our approach will improve the effectiveness of local search, and increase the ranking of local businesses. We introduced a new approach for enhancing local search engines efficiency in returning more accurate results. Our approach consists of semantically enriching the results of a query using LOD web content. This approach retrieves the comment tag of rdfs resources to find all relevant terms using stemming. Our preliminary evaluation demonstrated a better search with more accurate results.

6. FUTURE WORK

In pursuit of this approach, we will evaluate large datasets from the local business services. Furthermore, we will create a Plug-in application that can be embedded within local search engines. Finally, we are committed to optimizing the algorithm by improving its Precision and Recall.

7. REFERENCES

[1] Alistair M., Matthews, B., Wilson, M., and Brickley, D. (2005) "SKOS core: simple knowledge organisation for the web." In International Conference on Dublin Core and Metadata Applications, pp. pp-3.

[2] Bertin, M., & Atanassova, I. (2012). Semantic Enrichment of Scientific Publications and Metadata: Citation Analysis Through Contextual and Cognitive Analysis. D-Lib Magazine, 18(7/8). http://doi.org/10.1045/july2012-bertin.

[3] Bontcheva, K., Aswani, N., Kieniewicz, J., Andrews, S., & Wallis, M. (2015). EnviLOD WP5: Quantitative Evaluation of LOD-based Semantic Enrichment on Environmental Science Literature. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.385.31&rep=rep1&type=pdf

[4] Bontcheva, K., Kieniewicz, J., Andrews, S., & Wallis, M. (2015). Semantic Enrichment and Search: A Case Study on Environmental Science Literature. D-Lib Magazine, 21(1), 1.

[5] De Faria Cordeiro, K., de Faria, F. F., de Oliveira Pereira, B., Freitas, A., Ribeiro, C. E., Freitas, J. V. V. B., … others. (2011). An approach for managing and semantically enriching the publication of Linked Open Governmental Data. In Proceedings of the 3rd workshop in applied

computing for electronic government (WCGE), SBBD (pp. 82–95). Retrieved from https://www4.serpro.gov.br/wcge2011/artigos/artigo-an_approach_for_managing_and_semantically_enriching_the_publication_of_linked_open_governmental_data.pdf

[6] De Mendonça, R. R., da Cruz, S. M. S., De La Cerda, J. F. S. M., Cavalcanti, M. C., Cordeiro, K. F., & Campos, M. L. M. (2013). LOP: Capturing and Linking Open Provenance on LOD Cycle. In Proceedings of the Fifth Workshop on Semantic Web Information Management (pp. 3:1–3:8). New York, NY, USA: ACM. http://doi.org/10.1145/2484712.2484715

[7] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014, June). The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations (pp. 55-60).

[8] McBride, B. "The resource description framework (RDF) and its vocabulary description language RDFS." In Handbook on ontologies, pp. 51-65. Springer Berlin Heidelberg, 2004.

[9] Rizzo G. and Troncy R., “Nerd: a framework for unifying named entity recognition and disambiguation extraction tools,” in Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2012, pp. 73–76.

[10] Ronzhin, S. V. (2015). Semantic enrichment of Volunteered Geographic Information using Linked Data: a use case scenario for disaster management. Retrieved from http://dspace.library.uu.nl/handle/1874/316224

[11] Wetz, P., Stern, H., Jakobitsch, J., & Pammer, V. (2012). Matching Linked Open Data Entities to Local Thesaurus Concepts. In I-SEMANTICS (Posters & Demos) (pp. 6–11). Retrieved from http://ceur-ws.org/Vol-932/paper2.pdf

[12] Wimalasuriya, D.C., and Dejing, D. "Using multiple ontologies in information extraction." In Proceedings of the 18th ACM conference on Information and knowledge management, pp. 235-244. ACM, 2009.

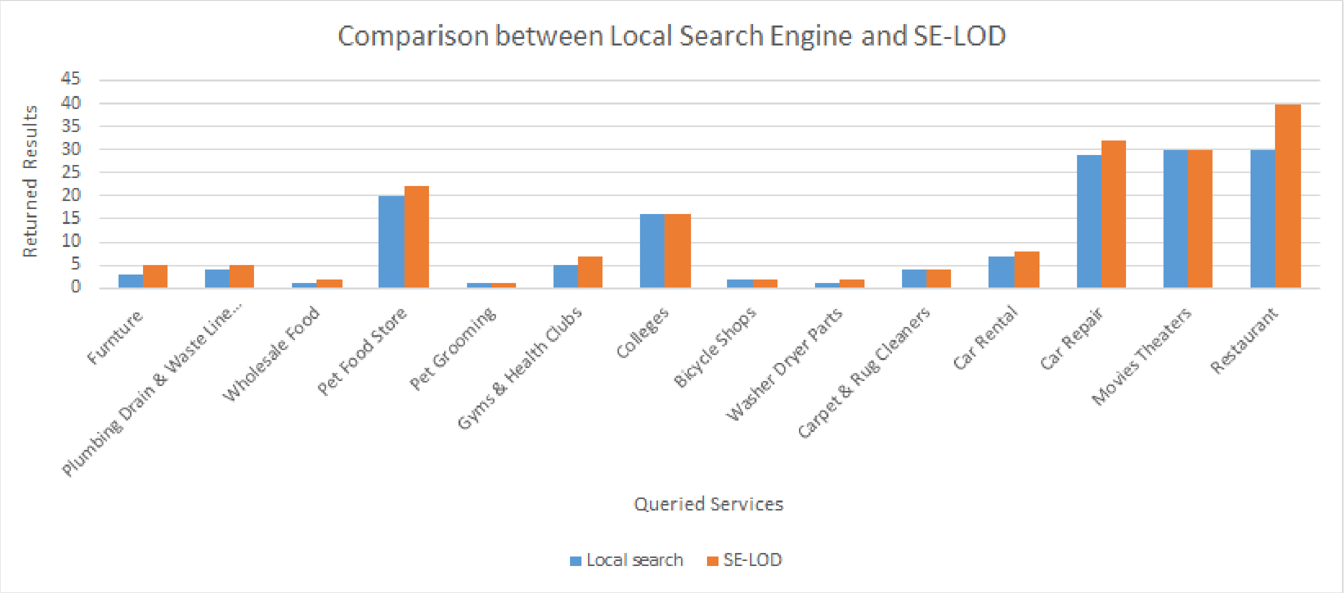


Figure 3: Comparison of Results between Local Search Engine and SE-LOD