Open Data for Local Search Challenges and Perspectives

[OD4LS 2016 Workshop overview]

Eric Charton
Yellow Pages Group
Montreal, QC, Canada
eric.charton@pj.ca

Nizar Ghoula
Yellow Pages Group
Montreal, QC, Canada
nizar.ghoula@pj.ca

Marie-Jean Meurs
Univ. of Quebec in Montreal
Montreal, QC, Canada
meurs.marie-jean@uqam.ca

ABSTRACT

Local search engines are specialized information retrieval systems enabling users to discover amenities and services in their neighbourhood. Developing a local search system still raises scientific questions, as well as very specific technical issues. Those issues come for example from the lack of information about local events and actors, or the specific form taken by the indexable data. Available open data can be exploited to dramatically improve the design of local search engines and their content.

The purpose of this workshop is to explore new fields of investigation both in terms of algorithmic approaches as well as originality of usable data. The workshop focuses on how open data can be used to enhance the capabilities of local search engines.

Keywords

ACM proceedings; Open Data for Local Search; Semantic Web; Local Search; Information Retrieval, Open Data

1. INTRODUCTION

Local search engines are specialized Information Retrieval (IR) systems enabling users to discover amenities and services in their neighbourhood (schools, businesses, hospitals, etc.).

Online local search is a growing field of economic activity explored by local business specialists such as Yellow Pages, Yelp or FourSquare, but also and increasingly, by major players in the web like Google, Yahoo or Facebook. It has been expressed as a key development axis for major companies like Facebook with its Local Search functionality, Yahoo with the ReachLocal platform, or Google with the My Business program. E-commerce specialists such as Amazon have also declared their interest for local search in the context of service offer, selling for instance services of local plumbers or lawyers.

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2890487 .

Online local related search engines and derivative products can be highly disruptive and innovative with propositions intended to translate the business of traditional companies dedicated to local search like YP USA, Solocal (Pages Jaunes) in France, or Yellow Pages in Canada. Those long-established actors, and their digital competitors face the challenge of developing new ways to integrate and publish exhaustive coverage of local businesses in their local search engines [6]. Over one century[1], companies specialized in printed telephone directories have developed methods and techniques to collect and structure information about local businesses in the perspective of publishing those information through a very specific medium: printed directories.

Nowadays, distributing local related information is mainly performed through modern and interactive platforms like smart-phones. Thus, this new way of listing businesses involves much more data. Those data are used to provide specific features: for example, automatically displaying propositions of related bars or restaurants when a user visits a shop implies collecting very precise latitudes and longitudes. Also, modern users equipped with computerized platforms want more accurate information about businesses: website URLs, pictures, video, rating, environment properties, and much more. These different types of data were not mandatory in printed directories, which were historically built from data feeds issued by phone companies, and simply made of company names, their physical addresses, and phone numbers. In a computerized context, this is far from adequate to build efficient systems. In such context this basic information must be supplemented by other resources. For instance for a merchant, one would look for a more descriptive textual content to help determine its business category (plumber, lawyer...), and extract metadata as hours of operation or specialized field of activity. All these mandatory contents are difficult to collect through traditional information extraction processes.

Due to those new needs in terms of data collection, indexation and retrieving, developing a local search system raises scientific questions, as well as very specific technical issues. One of the main challenges is the partial availability or even the absence of informative content related to local actors, merchants or service providers. As opposed to traditional search engines that benefit from the indexation of full docu-

[1]The name and concept of "yellow pages" came about in 1883, when a printer in Cheyenne, Wyoming, working on a regular telephone directory ran out of white paper and used yellow paper instead; see *en.wikipedia.org/wiki/Yellow_pages*

ments, this lack of content makes it hard to design efficient local search systems relying on vector space model.

For all of those reasons, the development of local search engines and related data architecture designs - including usage of open data - is an important topic of investigation for the World Wide Web community, and a highly active field of scientific research.

Some attempts were previously made to stimulate research in the field. For instance, the local search specialized company Yelp has successfully developed the Yelp Dataset Challenge[2]. In the context of the Dataset Challenge, Yelp released a specific dataset of local businesses that can be used to conduct experiments, and evaluate new algorithms.

The purpose of this workshop is to ask the scientific community how open data could be used in an innovative way for groundbreaking research to improve the design and relevance of local search engines. Thus, using semantic web datasets and natural language processing techniques, we want our workshop to present novel algorithms to improve geo-parsing [7], local business discovery, content extraction, event detection, semantic search, semantic relations, enrichment of listings, local search taxonomy building or enrichment, mining reviews and ratings, and any other topic potentially interesting for search engines supporting local search.

2. CHALLENGES

Using open data for local search is challenging, and opens new fields of investigation. Among the topics discussed during the workshop, we identify the following: the usage of open geographical data; the IR challenge; and the usage of open ontology in local search system.

In the context of IR applications, for instance, semantic web resources such as DBpedia contain keywords that could augment the content related to merchants and their categorization. An exhaustive ontology like DBPedia - but also like numerous others now available on the Linked Open Data (LOD) - could also be used for enriching ontologies associated with a local search service.

We believe that the geographicalal aspect of local search could find many opportunities for improvement using open data provided by cities or national organizations, such as descriptions of public institutions, opening hours, location of shopping centres, and other usable resources like points of interest (POIs).

We finally consider the algorithmic aspect of open data for local search as the most open and challenging one. The question here is to see if, outside all the established fields of IR research already involving studies about numerous dedicated algorithmic solutions, it is possible to find new innovative computing methods making use of open data to improve search engines.

In the following Sections, we explore in details each of those challenges, and show how they were addressed in the submissions.

2.1 Geographical Data Challenge

We identified the potential of open geographical data as one of the main potential contribution in local search [12]. Collecting geographical accurate information, and using it to identify or improve the data related to local actors is a major challenge.

[2]http://www.yelp.com/dataset_challenge

Collaborative data such as those made available by the OpenStreetMap Foundation can be of help to identify new dealers, and improve their geo-location. The same data can also be utilized to optimize the mapping systems of local search engines. Recently, numerous city like Toronto or Victoria in Canada, started providing open geographical data, like rooftop geo-locations, with a very high level of quality.

In this workshop we welcomed papers proposing solutions making use of open geographical data, and more than two third of the submissions make use of or are related to various sources of open geographical data.

2.2 Information Retrieval Challenge

A local search query is an IR action intended to find an entity given a specific location or a geographical area. This means that this type of search has two dimensions: (1) what is the user looking for, and (2) where is it geographically located. Compared to traditional IR actions, local search seems more complex since it tackles these two dimensions, the "what" contextualized by the "where".

The IR aspect of open data for local search can be covered in numerous ways. Keyword data, merchant categories, merchant properties (usually seen as facet search, deployed in most of local search engines) can be utilized to answers the "what" dimension. Regarding the "where" dimension, proper location of the merchants, open data for POIs, geographical entities that need to be specified, named and correctly positioned are critically needed, as well as ways to display the results using maps.

We welcomed any submission discussing this challenge and four accepted papers make use of open data to improve the relevance of results according to the "what" or the "where". Contributors proposed solutions to improve street search or ranking of search results, and relevance of search results, for example to find optimally geo-located houses. The keyword-based task was not covered: we did not receive any proposition addressing the important question of augmenting the recall in search results by using open data derived crawl or extracted keyword content. This is still a major challenge in local search, and we hope to see it more covered in future editions of OD4LS.

2.3 Ontological Challenge

Ontology-based IR has been the topic of multiple research works within the past decade [8, 9]. In an IR system like a search engine, ontologies can be used to improve the recall of documents with poor or insufficient content. This happens, for example, when the vector space model of an IR system is unable to match efficiently a query with a document, because of a lack of keyword content. Relying on ontologies to support search engines means facing many challenges, such as finding or defining the most suitable ontology for indexing and querying, using it to apply query expansion techniques [4], choosing the right knowledge representation, or extracting ontological relations [5].

To improve the recall of documents, the ontology used must properly represent the domain of the corpus to be indexed. In a complex IR system covering multiple domains like a local search engine, such an ontology may encompass multiple taxonomies organizing one or multiple subjects. This requirement for the ontology of an extensive and exhaustive domain coverage can be matched by the use of the semantic web resources [3], and more specifically the LOD

network and its content. LOD is made of open data sets structured as triples predicates fitting ontological definition, as for example DBpedia.

In this workshop we welcomed solutions combining approaches from new IR algorithms, exploitation of data collection [13], production or usage of knowledge graphs, data improvement using information extraction, or natural language processing techniques [16].

This field of investigation was not covered in any of the received submissions, though the knowledge representation utilized to build search results of local search engines can have a strong influence on the relevance of these results. Hence, we still consider that topic as one of the major sources of potential improvement for local search systems, and we hope to be able to cover it in future editions of the workshop.

3. OVERVIEW OF OD4LS TRACKS

The OD4LS workshop welcomed the submission of full papers (up to 6 pages), short papers (up to 4 pages) and posters/demos (up to 2 pages). All submissions were reviewed using a simple blind process, by at least three program committee members, and assessed based on their novelty, potential impact, and clarity of writing.

We welcomed papers from the industry, and academics. Presenting proprietary solutions was acceptable if they were making use of open data, and were sufficiently described to be reproducible. The experimental focus (experiments, metrics, and results) was not a key aspect for paper acceptance if the usage of open data was innovative enough.

All the accepted papers match these requirements, and the workshop proceedings describe creative solutions. Each of these solutions clearly presents the origin of the used open data along with the way they are accessed, empowering local search actors interested in these approaches with the capability to implement the described features.

According to the submitted propositions, we divided this first edition of the workshop in three tracks. The first one - *Content Enhancement* - is related to search engines using open content to improve their performances. The second track - *Algorithms Based on Open Data* - describes algorithmic methods specifically developed around open data. The third track is dedicated to demonstrations of algorithms, and solutions using open data.

3.1 Content Enhancement

For the first track of the workshop, we accepted three papers proposing novel approaches and experiments for content enrichment and categorization based on open data.

The first paper of this track is about **"Improving Local Search with Open Geographic Data"** by Chuankai An and Dan Rockmore [2]. Due to the lack of rich content for small businesses, open data directories can be used to enrich description of business units. The authors propose a method to support their intent by mining public datasets to find correlation between user preferences and several geographical features available from local data. The paper proposes a number of geographical features - number of reviews among cities, average distance between business units and other cities, density of business units, number of roads surrounding business units, location of business units on a road - to study how they influence the rating of businesses. Through some analyzes, the authors found that correlations exist between these features and users' rating and reviews.

The second paper is entitled **"A simple tags categorization framework using spatial coverage to discover geospatial semantics"** by Camille Tardy, Laurent Moccozet, and Gilles Falquet [17]. The paper proposes an approach to increase the geo-spatial recall for VGI (Volunteered Geographic Information) services based on spatial coverage, and a model for tag categorization. The authors support the usage of their model by describing an example for capturing the semantics of places to allow the disambiguation of tags. The model is based on the differentiation between spatial coverage and spatial references from the tags associated with the documents. It relies on the usage of external resources and multiple semantic services such as Geonames, OpenStreetMap or Wordnet.

This paper addresses two of the stated challenges: (1) the availability of geographical open data services covering the geospatial aspects - in this context, Geonames and OpenStreetMap -, and (2) the richness of terminological resources - in this context, WordNet - to identify the types and the semantic relations of entities in documents.

The third paper is about **"Semantic Enrichment for Local Search Engine using Linked Open Data"** by Mazen Alobaidi, Khalid Mahmood, and Susan Sabra [1]. The authors address the problem of incomplete listings of local services by proposing a new way for using LOD in semantic enrichment of query in local search engines.

The proposed system is based on four components: (1) a natural language processing module that uses the Stanford CoreNLP toolkit [11] to extract rich content to be used for (2) named entity recognition and disambiguation in order to extract named entities, and tag them with the proper semantic category, then (3) an entity URI look-up module that fetches the URI of given entities in order to match them with their representation in LOD, and finally (4) a triples extraction module that extracts triples from LOD about each entity, such as comments and linked categories, in order to enrich the content of the business description.

3.2 Algorithms Based on Open Data

For the second track of the workshop, we accepted two papers proposing novel approaches and experiments for algorithms and systems based on open data for a better user experience.

The first paper of this track is about **"Open Data Business Licence Usage to Improve Local Search Engine Content Ranking"** by Bob Lytle [10]. The author assesses the use of open data business license in order to provide end users with relevant and reliable authorized service providers, and reduce or remove providers with an invalid licence status, toward a better user experience.

The hypothesis is supported by a set of experiments that have been described and analyzed to address the issue of relevance for end users, which is not always the first criteria to be considered by local providers. This paper addresses multiple challenges about local search such as the gap between the relevance of the results (user experience), and the business model of local search providers (paid listings).

The second paper presents **"Lookupia : An Intelligent Real Estate Search Engine for Finding Houses Optimally Geolocated to Reach Points of Interest"** by Jonathan Milot, Patrick Munroe, Éric Beaudry, François Grondin, and Guillaume Bourdeau [14]. This contribution describes an intelligent real estate search engine. It makes

use of data from OpenStreetMap as primary source to determine the best property for a given set of users.

The search engine, called Lookupia, takes into account the proximity of the property from POIs specified by the user. The calculation of the proximity is based on the time needed to reach a POI, and weighted by the frequency of the visits the user intends to make to the POI. The result of the search is a number of properties, which are optimally geolocated with regard to the POIs. The search engine uses its own shortest path algorithm, adapted from Dijkstra's algorithm.

3.3 Demonstrations

In the demonstration track, we accepted a paper entitled **"Address Geocoding using Street Profiles for Local Search"** by Michael Peterman, Omar Benomar, Hacene Mechedou and Felix-Herve Bachand [15]. The authors address the issue of correctly geo-coding the position of a local business based on its address. The proposed method relies on a free and open government street lines data source. The address-geocoder transforms a street address into a location, typically measured in latitude-longitude coordinates. The address-geocoder is used in a search engine to relate spatial data to search results, and improve accuracy.

4. CONCLUSION

Organizing a new Workshop is always a challenge, especially in the context of a new and rarely investigated topic like the one we chose. We were hence very happy to receive more than enough propositions from both industrial and academic fields to build an attractive program. Our contributors make innovative use of multiple and diverse sources of open data: we learned a lot from their propositions, and discovered many sources of open data that deserve to be promoted. For this, we want to thank them for their efforts.

All the propositions selected for this first edition of the OD4LS workshop describe innovative solutions. This shows that there is a growing community actively involved in the exploration of open data usage for local search.

We found this emergence of interests encouraging and very promising at a time when local governmental organizations like city councils, departmental administrations or provinces deploy significant efforts to publish open data in the perspective of stimulating new economic fields.

After a promising start, we work now in the perspective of the second edition of OD4LS where we will try to extend the covered topics.

5. REFERENCES

[1] M. Alobaidi, K. Mahmood, and S. Sabra. Semantic Enrichment for Local Search Engine using Linked Open Data. In *OD4LS 2016 Workshop, WWW2016*. ACM, 2016.

[2] C. An and D. Rockmore. Improving Local Search with Open Geographic Data. In *OD4LS 2016 Workshop, WWW2016*. ACM, 2016.

[3] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story so far. *Int. jour. on semantic web and information systems*, 5(3):1–22, 2009.

[4] A. Bouchoucha, X. Liu, and J.-Y. Nie. Integrating Multiple Resources for Diversified Query Expansion. In *Proc. of the 36th Eur. Conf. on Information Retrieval, ECIR*, page 6. Springer, 2014.

[5] S. Brin. Extracting Patterns and Relations from the World Wide Web. In *The World Wide Web and Databases*, pages 172–183. Springer, 1999.

[6] M. Himmelstein. Local Search: The Internet is the Yellow Pages. *Computer*, (2):26–34, 2005.

[7] L. Hu, A. Sun, and Y. Liu. Your Neighbors Affect Your Ratings: On Geographical Neighborhood Influence to Rating Prediction. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 345–354, New York, NY, USA, 2014. ACM.

[8] V. Jain and M. Singh. Ontology-based Information Retrieval in Semantic Web: A Survey. *Int. Jour. of Inf. Tech. and Computer Science (IJITCS)*, page 62, 2013.

[9] S. Kara, Ö. Alan, O. Sabuncu, S. Akpınar, N. K. Cicekli, and F. N. Alpaslan. An Ontology-based Retrieval System Using Semantic Indexing. *Information Systems*, pages 294–305, 2012.

[10] B. Lytle. Open Data Business Licence Usage to Improve Local Search Engine Content Ranking. In *OD4LS 2016 Workshop, WWW2016*. ACM, 2016.

[11] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014.

[12] A. Markowetz, Y.-Y. Chen, T. Suel, X. Long, and B. Seeger. Design and Implementation of a Geographic Search Engine. In *WebDB*, volume 5, pages 19–24, 2005.

[13] G. McKenzie, K. Janowicz, and B. Adams. Weighted Multi-attribute Matching of User-generated Points of Interest. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 440–443. ACM, 2013.

[14] J. Milot, P. Munroe, E. Beaudry, F. Grondin, and G. Bourdeau. Lookupia : An Intelligent Real Estate Search Engine for Finding Houses Optimally Geolocated to Reach Points of Interest. In *OD4LS 2016 Workshop, WWW2016*. ACM, 2016.

[15] M. Peterman, O. Benomar, H. Mechedou, and F.-H. Bachand. Address Geocoding using Street Profiles for Local Search. In *OD4LS 2016 Workshop, WWW2016*. ACM, 2016.

[16] Y. Shin, Y. Ahn, H. Kim, and S.-g. Lee. Exploiting Synonymy to Measure Semantic Similarity of Sentences. In *Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication*, page 40. ACM, 2015.

[17] C. Tardy, L. Moccozet, and G. Falquet. A simple tags categorization framework using spatial coverage to discover geospatial semantics. In *OD4LS 2016 Workshop, WWW2016*. ACM, 2016.