

Transforming Wikipedia into a Search Engine for Local Experts

Gregory Grefenstette, Karima Rafes

▶ **To cite this version:**

Gregory Grefenstette, Karima Rafes. Transforming Wikipedia into a Search Engine for Local Experts. 2015. hal-01224114v1

HAL Id: hal-01224114

https://inria.hal.science/hal-01224114v1

Preprint submitted on 4 Nov 2015 (v1), last revised 30 May 2016 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Transforming Wikipedia into a Search Engine for Local Experts

Gregory Grefenstette
Inria Saclay/TAO, Rue Noetzlin - Bât 660
91190 Gif sur Yvette, France
gregory.grefenstette@inria.fr

Karima Rafes
Inria Saclay/TAO, Rue Noetzlin - Bât 660
91190 Gif sur Yvette, France
Karima.rafes@inria.fr

ABSTRACT

Finding experts for a given problem is recognized as a difficult task. Even when a taxonomy of subject expertise exists, and is associated with a group of experts, it can be hard to exploit by users who have not internalized the taxonomy. Here we present a method for both attaching experts to a domain ontology, and hiding this fact from the end user looking for an expert. By linking Wikipedia to this same pivot ontology, we describe how a user can browse Wikipedia, as they normally do to search for information, and use this browsing behavior to find experts. Experts are characterized by their textual productions (webpages, publications, reports), and these textual productions are attached to concepts in the pivot ontology. When the user finds the Wikipedia page characterizing their need, a list of experts is displayed. In this way we transform Wikipedia into a search engine for experts.

1. Introduction

In large organizations, such as multinational corporations, universities, and even large research centers, it can be difficult to know who is an expert about a given subject. A common response to this problem is to create an ontology of expertise and manually or automatically assign experts labels from this ontology. Beyond the cost or effort needed to produce the ontology, this solution creates an additional problem. Once such a knowledge base of experts exists, the searcher still has to internalize the ontology labels and their meaning in order to find the expert. The difficulty the user faces explains why some expert knowledge bases are found useful for one division in a large organization but be useless for another division which does not share the same terminology or perspective (Hahn & Subrami, 2000). We propose a method for finding experts that hides the pivot ontology from the user, and allows the searcher to browse Wikipedia, a resource that he or she is probably familiar with, in order to find his or her local expert.

2. Solution

Our solution involves mapping experts from a givein organization onto a domain ontology for their expertise, and then mapping Wikipedia articles into the same domain ontology. To find an expert on a certain field, the user searches the subject of interest in Wikipedia. A tab added to the Wikipedia page can be clicked to reveal a list of experts concerning the some topic mentioned in the page.

2.1 Example

Before explaining the mechanisms and language resources involved, let us see an example. In this example, our local experts are any of the research teams in the French nation-wide computer science public research institute, Inria[1]. The Inria Institute employs 3600 scientists spread over 200 research teams, each specializing in some branch of computer science and mathematics. In our example, finding an expert will mean finding a team who is competent to answer questions about a given subject.



Figure 1 An extra tab appears while browsing Wikipedia once the expert finder module is activated by a logged in user. Here the tab is labeled 'Inria' since we are searching for experts inside the Inria Institute.

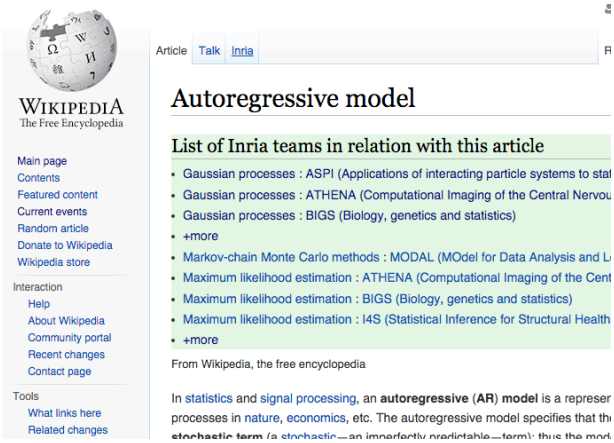


Figure 2. Clicking on the tab expands a box that shows the Inria teams that are associated with subjects on the Wikipedia page. For example, one subject mentioned on the page is "Gaussian processes" and 3 Inria teams that work in this domain are listed ASPI, ATHENA, and BIGS, with their expanded team names. Clicking on "Gaussian processes" goes to an ACM 2012 ontology page for this concept. Clicking on a team name goes to a web page from their 2014 annual report where a project involves "Gaussian processes". On this page, the user can find team members who are experts in the area. (See Figures 3 and 4)

In our solution, when someone is looking for an expert in a given subject inside the Inria institute, an additional tab appears to the user on the Wikipedia interface[2]. In Figure 1,

[1] https://en.wikipedia.org/wiki/Inria

[2] This tab appears when the user is logged in, and has the resource module for expert finding, this tab appears while the user browses http://en.wikipedia.org. Wiki resource

the tab appears with the label "Inria" to the right of the "Article" and "Talk" tabs above the article title. Clicking on the tab expands a box listing the ACM subjects found in the articles and the Inria research teams treating those subjects.

Both subjects and teams (see Figure 2) are linked to pages outside Wikipedia, so that Wikipedia has become a search engine with the user browsing towards their query (here "Autoregressive Models", with the pull down expert box corresponding the search engine results page, leading to outside content. The user can find the Wikipedia article closest to his or her concern, and use the expert finding tab to find local experts who know about the subjects on the page. This seems to us a natural and intuitive method for finding experts that obviates the need for learning the ontology by which the experts are connected to the topic page. Even if the connecting ontology terms are explicitly displayed in the results, the user need not ever use them in an explicit query.

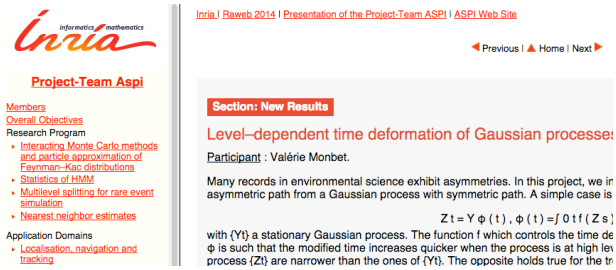


Figure 3. The Inria team annual report page found by following the link in the expert finding module. There the user sees that "Valerie Moribet" is involved in a project using Gaussian processes, and would probably be a good expert contact for finding out about Auto-regressive models.

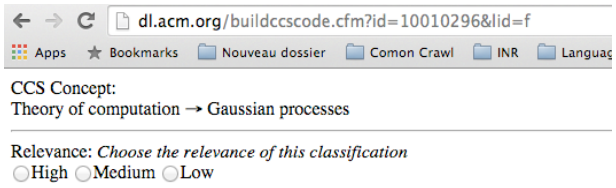


Figure 4. The ACM 2012 is the latest version of a computer science ontology created by the Association for Computing Machinery. "Gaussian processes" is a hyponym of "Theory of Computation". This specific concepts provides a link between the Wikipedia web page on "Autoregressive models" and the Inria team experts, but the searcher need not understand the ACM hierarchy nor know its contents in order to make the connection.

2.2 Underlying Language Resources

The connections between Wikipedia content and local experts pass through a shared pivot ontology. Both Wikipedia page content and the expert profiles are mapped into the same ontology categories which provide a pivot, or link between them. In our implemented example, we used the ACM 2012 Classification schema[3] as the shared ontological space. Here is an entry in this ontology:

<skos:Concept rdf:about="#10010296" xml:lang="en">
<skos:prefLabel xml:lang="en">Gaussian processes</skos:prefLabel>
<skos:altLabel xml:lang="en">gaussian process</skos:altLabel>
<skos:inScheme rdf:resource="http://totem.semedica.com/taxonomy/The ACM Computing Classification System (CCS)"/>
<skos:broader rdf:resource="#10010075"/>

This entry gives a synonym for "Gaussian processes", an internal ACM code for this concept (10010296), and a link to a hypernym (10010075), "Theory of computation". This SKOS taxonomy, augmented by any Wikipedia redirects as additional synonyms, was converted into a natural language processing program that associates the internal ACM code to raw English text (delivered as resource with this paper).

Wikipedia text was extracted from a recent dump.[4] The text was extracted from each article, tokenized, lowercased, and transformed into sentences. Each sentence was passed through NLP program associating ACM codes to each article. Any ACM concept appearing in more than 10,000 articles was eliminated as too general[5].

The source for expert profiles in our example were the public web pages of Inria teams 2014 activity reports[6]. Each page was downloaded, boilerplate removed, text extracted, tokenized lowercased and split into sentences, as with the Wikipedia text.

In all 3123 Inria web pages were associated 129,499 Wikipedia articles were tagged one or more of with 1049 different ACM codes.

2.3 Variations

Instead of using annual reports to create expert profiles, one could instead use the publications of researchers from a given team or research center.

In place the ACM hierarchy, one could extract a taxonomy of Wikipedia categories and subcategories[7].

For example, one could use MeSH as the anchor ontology and publications of doctors at local hospital to transform Wikipedia into a search engine of specialists for medical problems.

3. Related Work

Demartini (2007) proposed using Wikipedia to find experts by extracting expertise from Wikipedia describing people, or from Wiki editors that edited page corresponding to a given topic. West et al. (2012) tried to characterize what makes Wikipedia experts. Our approach allows us to connect to people who do not appear in Wikipedia either as subjects or editors.

4. CONCLUSION

We have presented a system for finding local experts using Wikipedia. It is constructed using a pivot ontology, indexing Wikipedia pages, and some textual representation of the experts (web pages, reports, or publications). The pivot ontology must be represented as a natural language processing resource, i.e. a resource that a program can apply to natural language text, that is applied to both Wikipedia pages and the textual representation of experts, pushing them into the same space. Once this mapping is done, a Wikimedia resource loader dynamically makes the connection once the user click

---

the user browses http://en.wikipedia.org. Wiki resource modules are additions that anyone can develop and activate. https://www.mediawiki.org/wiki/ResourceLoader/Developing_with_ResourceLoader explains them further.

[3] http://www.acm.org/about/class/class/2012

[4] http://dumps.wikimedia.org/biwiki/latest

[5] For example, ACM has a concept 'Women' under 'Gender'. 'Women' was eliminated while 'Gender' was retained.

[6] https://raweb.inria.fr/rapportsactivite/RA2014/index.html

[7] Starting from, for example, https://en.wikipedia.org/wiki/Category:Computer_science

on the expert fining tab, so neither the text of Wikipedia, not the expert profile text need be altered. An additional advantage of this system is that the user seeking an expert does not interact explicitly with the pivot ontology, but only with Wikipedia and the original textual representations of the experts. As the general public becomes more used to using Wikipedia to find information, they are able to find the best page that characterizes their need. And Wikipedia is transformed into a local expert search engine.

5. References

Demartini, Gianluca. "Finding Experts Using Wikipedia." FEWS 290: 33-41. 2007.

Hristoskova, Anna, Elena Tsiporkova, Tom Tourw, Simon Buelens, Mattias Putman, and Filip De Turck. "Identifying experts through a framework for knowledge extraction from public online sources." In 12th Dutch-Belgian Information Retrieval Workshop (DIR 2012), Ghent, Belgium, pp. 19-22. 2012.

Hahn, Jungpil, and Mani R. Subramani. "A framework of knowledge management systems: issues and challenges for theory and practice." In Proceedings of the twenty first international conference on Information systems, pp. 302-312. Association for Information Systems, 2000.

West, Robert, Ingmar Weber, and Carlos Castillo. "A data-driven sketch of Wikipedia editors." In Proceedings of the 21st international conference companion on World Wide Web, pp. 631-632. ACM, 2012.