The semantics of involvement: mining the web to locate biomedical expertise

Jorge Garcia-Flores*[1], Mark Deckert[2] , William Turner*[1] and Martine Hurault-Plantet[1]

[1]CNRS, LIMSI Orsay, F-91403 France
[2]UC Santa Cruz 1156 High Street, Santa Cruz, CA 95064

Email: Jorge Garcia-Flores*- jorge.garcia-flores@limsi.fr; Mark Deckert - mdeckert@ucsc.edu; William Turner*- william.turner@limsi.fr; Martine Hurault-Plantet - martine.hurault-plantet@limsi.fr;

*Corresponding author

Abstract

Background: In order to generate Google search strategies, we used NLP techniques and two separate corpuses to build semantic markers for identifying Argentinean scientists abroad. The first corpus was extracted from the Web of Science (WoS): using the bibliographical records of this database we identified the co-authors of papers, their institutional addresses, their subject area and country of residence. A second corpus of 50 documents was manually built in order to identify keywords and expressions for detecting people's biographical information on the Web.

Results: Using queries built exclusively from bibliographic data, 74% of the scientists in our test population were found, at an average cost of 21.3 search engine queries per scientist. Using biographic keywords and expressions, we got better recall (83%), but at a higher query cost (an average of 252.2 queries per scientist).

Conclusions: This paper reports on work aimed at constantly improving the cost efficiency of combining bibliographic data with biographic keywords when searching for scientist abroad on the Web. A combination of these two type of data seems the best compromise between a maximum scientist recall at a minumum query cost.

Background

The "semantics of involvement" question lies at the heart of the "brain drain/brain gain" debate in the field of migration studies. Field studies show that when scientists are abroad they group together into what has been called "Diaspora Knowledge Networks" (DKN) [1]. These self-organizing social structures serve, under certain conditions, to strengthen the science system in their countries of origin. While brain drain is undeniable, the existence of Diaspora Knowledge Networks in foreign countries generates positive externalities for countries of origin. Governments are actively seeking to strengthen these externalities through policies aimed at engaging DKNs in brain gain strategies. But how do governments reach out to people on the move, who often reside beyond national borders and who have not only the incentive but also the skills and qualifications to "make their life" abroad? Data on this population are available through a great many sources such as population censuses, labour force

surveys, administrative data and case studies. But each of these sources has its limitation: for example, population censuses are carried out too infrequently to capture the comings and goings of scientists on the move; labour force surveys are generally restricted to small sample sizes; national administrations provide work permits and temporary visas but generally don't collect and organize their data using the concepts, definitions and classifications needed for calculating international statistics; case studies are often one shot ventures which are too costly to update on a regular basis. The question, then, is to know if the Web can be used to overcome these different limitations.

Methods

We proceeded in three steps. First, we manually identified 23 Argentinean born scientists working abroad, downloaded the WoS record of their most recent publication, and built a corpus containing their diasporic evidence (CVs, personal pages, mentions in organizational charts or professional networks) published on the Web. We next used NLTK pos-tagger [2] to extract nouns, noun phrases, cities and insitutions from the WoS records. Our hypothesis was that by combining an author's name with semantic markers concerning her institutional address, her country of residence and the noun terms extracted from the title of her publication we could produce a set of queries that would bring diasporic evidence to the top 10 results of Google. In order to test this hypothesis, we experimented different semantic marker combinations, and compared the results with those obtained by using a list of keywords and expressions for directly identifying biographical data.

Results

A total of 24558 queries were produced by a) combining WoS bibliographic data for each scientist and b) combining the scientist's name with biographic keywords. A "hit" occurred when the appropriate diasporic evidence (CV or Web page) in our biographic corpus was found in the top ten Google results returned for a query. Queries were sent to Google by a Perl script that automatically registered hits. Evaluation was performed by calculating $recall(h) = \frac{|D_h|}{|S|}$, where D_h is the set of hits obtained by using a specific combination of semantic markers h, and S the set of all the scientist in the corpus; $Brecall(h) = \frac{|D_h - D_b|}{|S - D_b|}$, where D_b is the set of distinct hits found by the baseline query b; and $\bar{q}(h) = \frac{|Q_h|}{|D_h|}$, where $\bar{q}(h)$ is the average number of queries necessary to find one scientist with heuristics h. This metric represents the cost (in queries) of specific levels of recall. Using queries built exclusively from bibliographic data, 74% of the scientists in our test population were found (17 out of 23), at an average cost of 21.3 search engine queries per scientist. Using biographic keywords and expressions, we got better recall (83%), but at a higher query cost (an average of 252.2 queries per scientist). The optimum heuristic (the set of markers that allow the maximum amount of evidence with a minimum amount of queries) is a combination of bibliographic data and biographic keywords.

Conclusion

Our work is aimed at systematically updating our understanding of the impact of scientific mobility on national science systems. We've focused in this paper on building a link from bibliographical data downloaded weekly from the Web of Science to the diasporic evidence (CVs and personal Web pages) of the authors. The next step will be to use this career information to automatically class authors as being a "home country scientist" (has trained and worked only in the country of origin); a "circular scientist" (has trained abroad but has elected to come back to work at home); or a "diaspora scientist" (has received initial training at home but has elected to reside abroad). IE Techniques from Web People Search (WePS) [3] could be used to class scientist in one of these categories. Case studies in the migration field have shown that these categories designate actors who, through their networks, are able to mobilize different configurations of human, institutional and cognitive resources for home country development. We are trying to develop tools for mining the Web in order to test this hypothesis.

References

1. Turner WA, Meyer JB, de Guchteneire P, Azizi A: *Migration, Internet und Politik. Potentiale für Partizipation, Kommunikation und Integration*, Wiesbaden: VS Verlag 2009 chap. Diaspora Knowledge Networks.
2. Bird S, Klein E, Loper E: *Natural Language Processing* 2008, [http://www.nltk.org/book-1].
3. Sekine S, Artiles J: **WePS2 Attribute Extraction Task**. In *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference* 2009.