Published in: Proceedings of the 5th International Conference on Knowledge Discovery and Information Retrieval (KDIR), 2013.

Challenges and Potentials for Keyword Extraction from Company Websites for the Development of Regional Knowledge Maps

Christian Wartena[1] and Montserrat Garcia Alsina[2]
[1]*Hochschule Hannover, Expo Plaza 12, 30539 Hannover, Germany*
[2]*Universitat Oberta de Catalunya, Rambla del Poblenou, 256, 08018 Barcelona, Spain*
christian.wartena@hs-hannover.de, mgarciaals@uoc.edu

Keywords: Text Mining, Keyword Extraction, Territorial Intelligence, Regional Innovation Systems.

Abstract: Regional Innovation Systems describe the relations between actors, structures and infrastructures in a region in order to stimulate innovation and regional development. For these systems the collection and organization of information is crucial. In the present paper we investigate the possibilities to extract information from websites of companies. First we describe regional innovation systems and the information types that are necessary to create them. Then we discuss the possibilities of text mining and keyword extraction techniques to extract this information from company websites. Finally, we describe a small scale experiment in which keywords related to economic sectors and commodities are extracted from the websites of over 200 companies. This experiment shows what the main challenges are for information extraction from websites for regional innovation systems.

1 INTRODUCTION

A basic prerequisite for regional development, for stimulating regional innovation and for improved cooperation between companies in a region, is a high quality overview of all companies in that region, their activities and their strengths. However, hardly any such overviews exist. Information about companies is often incomplete, outdated or focuses only on one specific branch of industry. Official lists, if available at all, also suffer from the problem that officially registered companies might be inactive or just be administrative constructs being part of a larger company, that has to be considered as a unit when activities, customers, products, services, etc. have to be described.

Building a regional overview we face several challenges. In the first place, we have to identify which companies in a region are involved in innovation and regional development. The second challenge is the selection of information sources. In the third place, we should determine which information from each company is relevant to identify its strengths, and to promote innovation and regional development. Finally, we have to describe the information about a company, its activities, products and customers in a uniform way. As with regard to second challenge, it seems natural to extract information from the company's website to obtain basic information (formal data, products and activities) about a company, since virtually every company has a website. There are other sources of up to date information, like patent databases or commercial directories. However, in the present paper we will investigate the possibilities to use the websites of companies, leaving other sources for future work. We identify what the main challenges are and we explore the possibilities to use them as an information source to describe the knowledge existing in a region. In order to approach the third and fourth challenge we combine expertise from competitive and territorial intelligence and data mining and more specifically keyword extraction (KWE).

While we sketch the overall approach for the construction of comprehensive regional knowledge maps in (Garcia-Alsina et al., 2013), in this paper we focus on the possibilities of KWE. In a small empirical study we investigate the possibilities to extract knowledge from company websites for the construction of regional knowledge maps of innovation systems. We use simple thesaurus based KWE techniques, but show how these techniques can provide very rich information about companies when used in combination with a highly structured thesaurus. In contrast to previous work on KWE we do not obtain a flat list of keywords but we get a keyword based description of a company for different aspects. In the current study we have evaluated the extracted keywords for two aspects: the economic sector a company belongs to and products produced by the com-

pany. In the experiment we also see the limits of the approach. In the first place, the extraction of keywords from complete websites consisting of pages on completely different topics, differs substantially from classical KWE of uniform texts. Furthermore, we do not try to find keywords for the texts, but keywords that apply to the company behind these texts.

The remainder of the paper is organized as follows. In section 2 we discuss the theoretical background and related work in territorial intelligence and text mining. In section 3 we describe an experiment to access the feasibility of KWE as a tool to support the construction of regional knowledge maps. The results of the experiment are given and discussed in section 4. We finish the paper with a conclusion and an outlook to future work.

2 RELATED WORK

2.1 Territorial Intelligence

Studies about the regional economics are done with methods coming from different disciplines: Economy, Geography, Sociology, Science of Education, Information and Communication Science (Mollo, 2009). More specifically, in the last 15 years a line of research has been developed, that identifies innovation processes as a key to the regional development (Girardot and Brunau, 2010; OECD and EUROSTAT, 2005).

Territorial Intelligence (TI) is a collective posture that explores the territorial possibilities by the collection of information and its treatment in order to anticipate risks and threats (Herbaux, 2008). To (Girardot, 2008) the territorial intelligence is the science whose object is the sustainable development of territories and whose subject is the territorial community. TI focuses not only on the economic efficacy and efficiency of the development model, but also includes all the dimensions that affect sustainable development (social, political, cultural, and environmental). It involves information about resources available in the territory, products, services (individual or collective, private or public) people needs, which activities are taking part in a territory and the territorial dynamics (Girardot and Brunau, 2010; Girardot, 2008; Herbaux, 2008). Hence, the collection of information about the territory plays an important role to planning strategic actions to put in touch different actors.

According to (Jimenez et al., 2011; Asheim and Gertler, 2005; Doloreux and Parto, 2005; Lundvall and Johnson, 1994) a Regional Innovation System (RIS) or National Innovation System (NIS) consists of the relation between actors, structures and infrastructures involved in a region, as well as the knowledge flow between them, that serve the goal of bringing innovative performance of companies in the region. Besides, an efficient system of distribution and access to knowledge contributes to increase the amount of innovative opportunities (Lundvall, 1998; David and Foray, 1995). More specifically, RIS and NIS as research area have developed a framework to study the factors that enable innovation and regional economic development. Some of the features of this framework are: a) institutions (rules, norms, or organizations) (Lundvall and Christensen, 2003; Nelson, 1993), b) innovation processes, c) knowledge's flows which take place within the regional learning process (Lundvall, 1992); d) the social capital and the regional context in which innovation happens, and that provides a set of rules, conventions and norms that prescribe behavioral roles and shape expectations (Doloreux and Parto, 2005; Nahapiet and Ghoshal, 1998); e) influence of national or local or social idiosyncrasies, which influence the social process (Salavisa and Vali, 2012; Lundvall, 1998), and f) the intersectorial differences to explain innovation activities (Doloreux et al., 2008).

Considering these antecedents, we underline the relevance of the role of: a) knowledge in the innovation process and consequently, knowledge management in firms and regions, where knowledge maps play an important role (Barinani et al., 2011; Driessen et al., 2007), b) the role of companies as actors in a regional innovation system, and c) the information about the environment that the companies should achieve and d) controlled vocabularies, like thesauri, with a common understanding of central terms to classify and retrieve knowledge (Färber and Rettinger, 2013; Garcia-Alsina and Ortoll, 2012; Canongia, 2007; Escorsa et al., 2000). Taking into account the amount of documents, it is difficult extract relevant information, and classify this information, so semantic technologies and data mining are needed (Färber and Rettinger, 2013; Eckert et al., 2007).

2.2 Keyword Extraction

Our current goal is to classify companies and other organizations. A number of information sources can be used to do this. In the present study we focus however exclusively on the web sites as a source of information. Thus we can consider our task as a text classification problem. Text classification is a well studied field. A good overview of techniques and approaches is given by (Sebastiani, 2002).If there is a large number of categories, which is usual the case if each term

of a thesaurus is considered as a potential document class, the standard classification approach cannot be applied, as not enough training data are available. In these cases a KWE approach is usually better suited. In its basic form, KWE is nothing else than selecting the most salient terms of a document. Determining such terms has been studied since the mid of the previous century (Salton and Buckley, 1987). However, salience turns out not to be the only criterion for keywords. More features can be found, indicating whether a term is suited as a keyword or not. These features, together with the relevance weight of the term, can be used in a supervised machine learning setting to learn how to distinguish keywords from non-keywords. This approach to KWE was proposed by (Frank et al., 1999) and (Turney, 2000).

Alternatively, the set of possible keywords can be restricted by the terms of a thesaurus or some other restricted vocabulary. This approach is followed by (De Campos et al., 2007) who use information from the thesaurus in combination with Bayesian statistics to suggest keywords. (Wang et al., 2007) uses PageRank to determine the most central words in the graphs which is constructed with the WordNet relations between the potential keywords. The community structure formed by the potential keywords and their relations is used by (Grineva et al., 2009). In (Malaisé et al., 2007a) and (Gazendam et al., 2010) words with a large number of relations get higher weights in order to promote central concepts.

3 SYSTEM DESCRIPTION

In order to investigate the potentials of KWE we crawl the pages from the websites of over 200 companies and store them in a local repository. We annotate the data manually for several aspect in order to be able to evaluate KWE algorithms. In the following we will describe the thesaurus, the data sources and the KWE in more detail.

3.1 Thesaurus

We use the STW Thesaurus for Economics[1] as a source for potential keywords. The STW is organized in subthesauri. In order to describe the company we extract only words that are used as descriptors in the subthesauri *Commodities* and *Economic Sectors*. For classification of the economic sectors it might seem more natural to use the NACE (Nomenclature statistique des Activités économiques dans la Communauté Européenne.) classification, that is an official standard in the European Union for the classification of economic activities. However, the STW Economic Thesaurus has the advantages that

1. it is available in SKOS format (Isaac and Summers, 2009), and thus can be read easily by text analysis software;
2. it has a lot of different descriptors for each category, that might be found in texts;
3. it contains a classification of products and related areas as well.

Especially the second property is essential for our approach: since the descriptors are usually short terms, that are likely to be found in texts, we can use these descriptors as a list of potential keywords. In order to use the approach for the construction of a regional knowledge map, we either have to define a mapping from STW sectors to NACE categories, or we have to transform the NACE thesaurus into SKOS and populate the categories with appropriate descriptors.

The main disadvantage of the STW is, that it is designed to index documents about economics, not to classify companies and their products. We are not aware of any usage of the STW for this purpose. The difference between the two goals might mainly be one of focus and coverage: the sectors that is most written about, and thus are expected to have a good coverage in the STW, need not to be the same sectors to which many companies belong.

As already said above, the STW Thesaurus for Economics consists of 7 subthesauri. Each thesaurus has several subject categories, that can contain subcategories. Subject categories represent either a technical intermediate level of description (e.g. *Branches of Industry*) or classes of products, economic sectors, etc. Each category has a denotation and a number of descriptors. The descriptors are subclasses or examples of the category they belong to. E.g. *fashion* is a descriptor of the category *Textile and Clothing Industry* (W.06.01.11). Also the technical categories have descriptors, that usually refer to instances that cannot be classified into one of the subclasses.

3.2 Data Sources

In order to investigate the possibilities of describing companies by extracting keywords from their websites we have compiled a set of 229 companies from 10 economic sectors. These companies are not situated in one region, as we do not yet aim at the construction of a regional knowledge map in this phase of the development. All companies are German and

[1] http://zbw.eu/stw/versions/latest/about.en.html

have German websites. We have classified the companies and their products according to the main categories of the STW. The companies were selected by students, that had the task of selecting 20 websites of companies in a specific branch or sector as a source for keywords in the related domain.

We have used the crawler4j[2] to crawl the websites. For 21 companies the crawling was not successful and no pages could be retrieved. Thus 208 companies remain. Since a few companies have a site with a high number of pages, we limited the number of pages to be retrieved to 120. The limitation services the practical goal of keeping the size of the corpus moderate, but also has more fundamental reasons: A few companies have very large websites, which makes the amount of information per company very unbalanced. Moreover, we expect that even a large company should be described rather well on the first two levels of a web site. If we crawl in a breadth first way, as we do, we might expect that at some point we have seen the core information of a company. If more pages follow, we might get more and more specific information on detailed topics, that even could obscure the more important and central information. The limit of 120 is rather arbitrary and turned out to be a size that allows us for almost all companies in our list to crawl the complete site. In total 14 673 pages were retrieved, which averages to 70.5 pages per company.

We did not do any boiler plate removal since it turned out that in many cases essential information is removed. E.g. a list of products or departments is often given as a menu, that might be removed. For companies with very limited websites, or in cases where the products or departments mentioned in the list point to other domains (we crawl only pages from one domain for each company) the most important information then would be removed.

3.3 Establishment of a Ground Truth

In order to create a base for evaluation of the automatic KWE, we classified all companies manually with the subject categories of the subthesauri *Commodities* and *Economic Sectors*. These categories are quite broad and can be assigned with a low rate of error and subjectivity and do not require very deep analysis of the information available about a company. In many cases several economic sectors and products had to be assigned. In the first place this was necessary for a number of large companies or business groups that are active in several (related sectors), like machinery construction and electrical industry. The second reason for multiple assignments are businesses that can be viewed from different perspectives. E.g. there is a number of goat farms in our data set. These farms can be classified as *Animal Husbandry* (W.01.02), but since they usually produce goat cheese, they also can be classified as *Food and Tobacco Industry* (W.06.01.12), but as they usually have a shop where they sell the cheese, *Retail Trade* (W.10.05) is also not completely wrong. Moreover, such farms often offer possibilities to view the animals to the public and they have a small restaurant or even offer cottages for rent. Thus a number of further subject categories apply.

3.4 Keyword Extraction

As a first step for KWE all texts are analyzed using a GATE pipeline (Cunningham et al., 2002) that consists of a language guesser, a tokenizer, a sentence splitter, a part of speech tagger and lemmatizer an ontology lookup component and several JAPE grammars. The language guesser is used to ensure that only German texts are analyzed. The ontology lookup component Apolda (Wartena et al., 2007) finds all labels of thesaurus terms in the texts, including multiword terms and inflected forms. The JAPE grammar formalism is part of the GATE software, and enables the definition of patterns over words and previous annotations. The Jape Grammars are used to extract addresses, phone numbers and names of companies.

Since advanced algorithms for KWE usually have been reported to give only limited benefit over a simple tf.idf weighting scheme, we decided to extract the words with the highest tf.idf score. In a later phase other algorithms can be used, but using a simple algorithm makes identification of problems easier. As well for the computation of the document frequencies as for the computation of the term frequencies we consider the whole of all web pages of a company as one document. The intuition behind this is, that a term like *Geschäftsführer* (Manager) that is found on 348 pages of 111 companies in our data set is much less indicative for the business of a company than the word *Schiff* (ship) that occurs on approximately the same number of pages (330) but only for 29 companies (a number of shipyards is in our data set). The tf.idf value is computed as usual:

$$\text{tf.idf}(w,d) = n_d(w) \cdot \log\left(\frac{N}{\text{df}(w)}\right) \qquad (1)$$

where $n_d(w)$ is the number of occurrences of w in document d, N is the total number of documents and $\text{df}(w)$ is the number of documents in which w occurs.

[2]http://code.google.com/p/crawler4j/

Table 1: Precision of extracted keywords for economic sectors and commodities.

	Sectors	Commodities
prec@1	0.076	0.27
prec@2	0.088	0.27
prec@5	0.063	0.23

4 RESULTS AND DISCUSSION

All extracted keywords are descriptors in the subthesauri *Commodities* or *Economic Sectors* in the STW. As mentioned above, we classified the companies according to the main categories in these subthesauri. We consider a keyword correct if it is a descriptor of a category assigned to the company. Thus we can compute the precision, i.e. the fraction of correct keywords, for each top *n* elements of the ranked keyword list. The precision of the top *n* keywords is referred to as prec@n. Since we have no exhaustive list of keywords that should be assigned to a company, we cannot compute a recall value. The precision for economic sectors and commodities is given in Table 1.

The results show immediately that the extraction of product categories is much easier than the extraction of economic sectors. This is not very surprising since companies will tell about their products on their website, not about the sector they operate in. Even worse, they might write about the use of their products in the sectors their customers are from.

Though we did not use any advanced technique for KWE, but rather establish a baseline, the results are in the same order of magnitude as reported in literature. E.g. (Medelyan and Witten, 2005) find a prec@5 of 0.21 for assignment of terms from the Agrovoc thesaurus to documents, (Gazendam et al., 2009) find a prec@5 of 0.23 for assignment of terms from the Dutch GTAA thesaurus for audio-visual archives. In this context it should be noted, that the comparison is useful to get a feeling for the order of magnitude we have to think about. In classical work on KWE, keywords are used as general descriptors of a text. Here, we use keywords to describe a specific aspect of a text. Furthermore, we look for keywords, that are not necessarily good descriptors for the texts they are extracted from, but we look for keywords describing the company behind the website.

When looking at the results for the economic sectors, the approach seems not to be very successful: the majority of the keywords found where judged to be irrelevant in our evaluation. In the following we will, take a closer look, at the irrelevant keywords that were found and propose ways to improve the results.

4.1 Error Analysis

For most cases, in which no keywords from the right class have been assigned the underlying property is data sparseness. Sparseness might have led either to unprecise manual classification, or to lack of thesaurus terms that could be used by the analysis of the internet sites. A typical case of the former is a set of over 10 companies that are specialized in library (software) systems. The closest branch in the STW is *Information Services* (W.19.05). The KWE finds mainly words related to libraries. Since a software company writing software for libraries is not a library, these keywords are evaluated as incorrect. A related problem arises from our decision to use the subject categories to classify companies and to evaluate the descriptors found as keywords. In most cases this works quite well. E.g. the category *Fishery* (W.03) has descriptors like *Aquaculture*, *Fishery*, *Fishermen*, *Fishery fleet*, etc. These terms both are likely to occur in texts and are good descriptors for companies in this sector. However, a category like *Branches of Industry* (W.06.01) has subcategories, representing specific branches, that have descriptors describing them, but it has also descriptors itself. These descriptors represent branches, like *Toy industry* that do not have other descriptors themselves. Consequently we have to annotate a toy manufacturer as *Branches of Industry* (W.06.01), which is not very adequate.

If we analyze the wrong keywords, we find five brought classes of errors:

Minor Aspects. In a number of cases, the keyword found is in fact correct, but its category was not used to classify the company, usually because the category does not reflect the main business of that company. This is especially the case for the products category, where we usually assigned only the main category, while many other product categories might apply.

Random Classes. In some cases the extracted keywords seem completely random. This is the result of the idf-weighting scheme. If some seldom term is mentioned two or three times on a small website it will get a very high score. This effect is reinforced by the fact that the correct terms usually do not have a very high tf.idf value, since we started collecting several companies from the same sector. This problem is observed frequently in thesaurus based KWE. A method to exclude completely irrelevant terms, is to analyze the (thesaurus) relations between the keywords initially found: if a keyword has no relation (or only few relations) to other keywords, it is likely that it is not related to the main topic of the document. Var-

ious proposals have been made to operationalize this idea (Malaisé et al., 2007a), (Grineva et al., 2009), (Gazendam et al., 2010).

Customer Classes. Web pages are usually designed for the customers of a company. This might result in texts that contain much more terms describing the branch of the customers of a company than the company itself. E.g. for most pharmaceutical companies descriptors for the branch *Health care systems* (W.25), which is reserved for hospitals and physician practices, are found. Another typical example are the shipyards, classified as *Vehicle Construction* (W.06.01.03), but for which many terms are found from the categories *Port Management* (W.12.01.03.03) and *Shipping* (W.12.01.03).

Ambiguous Terms. Some terms are in some way ambiguous and might refer to two different branches of industry, but are only used for one of both in the STW. E.g. the word *radio* is used a descriptor for *Broadcasting Industry* (W.19.03), but is in our data set found for a company producing radio sets. In a number of cases, the structure of the thesaurus is also problematic. E.g. the term *sail*, sometimes found on the pages of shipyards, is an alternative term for *off-the-peg textiles* that belongs to the category *Clothing* (P.19).

Broader Terms. A typical error is the selection of a term that is too broad, but in fact not incorrect. This is a common problem for thesaurus based KWE and alternative evaluation methods have been proposed to deal with this (Medelyan and Witten, 2005) and (Gazendam et al., 2009).

Finally, keywords that do not match our self defined gold standard do not have to be wrong. Exactly these keywords might be very interesting and informative. The keywords that do not match with the obvious economic sector, might give hints to hidden expertise. E.g., a company from the sector *Vehicle construction* (W.06.01.03), which main product is ships construction (*Watercrafts*; P.09.03), but with an extracted keyword *Textiles* (P.18), might indeed have expertise on textiles that could also be useful for other companies. Nevertheless, first the main classification for each company has to be established.

4.2 Thesaurus Enhancement

The majority of the problems arises from data sparsity in the thesaurus. It makes of course no sense to add ad-hoc categories for products or sectors that are completely missing, like medical equipment industry, since we use a thesaurus because it is standardized and stable. However, we can add more descriptors, non-preferred labels, spelling variants and synonyms to the thesaurus. This would let the original thesaurus intact. The extended version of the thesaurus needs only to be used by the KWE algorithm, while the results are still descriptors or categories from the official thesaurus. When more labels are available, especially synonyms and alternative labels, the final result is based on much more data, and the occurrence of an irrelevant word will have less impact. Now, in many cases only two or three thesaurus terms are found on a page, and it is more or less by chance whether these are relevant or irrelevant terms. Enrichment of a thesaurus to enhance thesaurus based KWE is a common method used e.g. by (Tiun et al., 2001), (Malaisé et al., 2007b).

There are various ways to extend the thesaurus in the proposed way. In the first place, the STW has many links to equivalent terms in other thesauri, like the German Subject Headings Authority File (SWD), that has many alternative labels for each term. Using the link between the SWD and the STW these alternative labels could be added to the STW as well. Furthermore, general dictionaries could be used to find synonyms and finally there exist very good methods to find synonyms statistically in large corpora (see e.g. (Weeds et al., 2004) for an overview of methods of distributional similarity).

4.3 Future Work

For future work, we have to extract more types of information. Obviously, there are other sources of information and a challenge for the future will be to integrate the knowledge about the regional economy that comes from different sources, uses different vocabularies, is written for different target audiences and might even be contradictory.

The other line of research we have to follow is the improvement of the information extraction methods. Here we can start with more advanced methods, but we also have to take into account the peculiarities of websites, the large size of web sites, (consisting of 70 pages on average, while a lot of KWE research is done on short abstracts) and the fact that the keywords should not describe the website, but the company behind the website.

5 CONCLUSIONS

We have argued that automatic methods to find and aggregate information on the internet are deserved for the construction of concise regional knowledge maps

in which economic activities and potentials of companies and other organizations in a region are described. The abundance of information rich websites of companies seems to offer good possibilities to do so.

In an experimental study we have shown that websites of companies indeed can be used as a source of information. In the first place, factual information like addresses and phone numbers can be extracted relatively easily, which we did not discuss in more detail in the paper. Secondly, the description of various aspects of a company can be described separately by automatically extracted keywords when using an appropriately structured thesaurus. Thus we go a step further than is usually done in KWE, where only one flat list of keywords is produced. We have evaluated the keywords for two aspects. For the commodities this works quite well, yielding a quality of results comparable to KWE results reported in literature. For the aspect *economic sectors*, the precision of the results is not very high. Here we are confronted with the problem that websites do not describe companies but are in the first place description and advertisement for a wide audience and potential customers. Thus the terminology on a good website will tell us at least as much about the audience as about the company. The information about products is much more factual and less dependent on a point of view than the economic sector, which explains the difference in performance between the two aspects. For other aspects we will have to develop techniques that are better able to separate the information about the writer and the audience of a website.

Overall we have shown that analysis of web information and KWE constitute an interesting source of information for territorial intelligence and the construction of concise knowledge maps, that is worth to be explored further.

ACKNOWLEDGEMENTS

The research presented in this paper was partially funded by the Spanish Ministry of Education, Culture and Sport (Ref. CAS 12/00155).

REFERENCES

Asheim, B. and Gertler, M. (2005). The geography of innovation: regional innovation systems. In Fagerberg, J., Mowery, D., and Nelson, R., editors, *The Oxford Handbook of Innovation.*, pages 291 – 317. Oxford University Press, Oxford.

Barinani, A., Agard, B., and Beaudry, C. (2011). Competence maps using agglomerative hierarchical clustering. *Journal of Intelligence Manufacturing*, pages 1–12.

Canongia, C. (2007). Synergy between competitive intelligence (CI), knowledge management (KM) and technological foresight (TF) as a strategic model of prospecting — the use of biotechnology in the development of drugs against breast cancer. *Biotechnology Advances*, 25(1):57–74.

Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). A framework and graphical development environment for robust nlp tools and applications. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 168–175. ACL.

David, P. and Foray, D. (1995). Assessing and expanding the science and technology knowledge base. *STI Review*, 14:13–68.

De Campos, L. M., Fernandez-Luna, J. M., Huete, J. F., and Romero, A. E. (2007). Automatic indexing from a thesaurus using bayesian networks. In Mellouli, K., editor, *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 865–877. LNCS 4724, Springer.

Doloreux, D., Nabil, A., and Landry, R. (2008). Mapping regional and sectoral characteristics of knowledge-intensive business services: Evidence from the province of Quebec (Canada). *Growth and Change*, 39(3):464–496.

Doloreux, D. and Parto, S. (2005). Regional innovation systems: Current discourse and unresolved issues. *Technology in Society*, 27:133–153.

Driessen, S., Huijsen, W., and Grootveld, M. (2007). A framework for evaluating knowledge-mapping tools. *Journal of Knowledge Management*, 11(2):109 – 117.

Eckert, K., Stuckenschmidt, H., and Pfeffer, M. (2007). Interactive thesaurus assessment for automatic document annotation. In *Proceedings of the 4th international conference on Knowledge capture.*, pages 103–110. ACM.

Escorsa, P., Rodriguez, M., and Maspons, R. (2000). Technology mapping, business strategy and market opportunities. *Competitive Intelligence Review*, 11(1):46–57.

Färber, M. and Rettinger, A. (2013). A semantic wiki for novelty search on documents. In *Proceedings of the 13th Dutch-Belgian Workshop on Information Retrieval*, pages 60–61, Delft.

Frank, E., Paynter, G. W., Witten, I. H., Gutwin, C., and Nevill-Manning, C. G. (1999). Domain-specific keyphrase extraction. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, IJCAI 99, Stockholm, Sweden, July 31 - August 6, 1999.*, pages 668–673.

Garcia-Alsina, M. and Ortoll, E. (2012). *La Inteligencia Competitiva: evolución histórica y fundamentos teóricos*. Trea, Gijón.

Garcia-Alsina, M., Wartena, C., and Lieberam-Schmidt, S. (2013). Regional knowledge maps: potentials and challenges. In *Fifth International Conference on Knowledge Management and Information Sharing (KMIS 2013)*.

Gazendam, L., Wartena, C., and Brussee, R. (2010). Thesaurus based term ranking for keyword extraction. In Tjoa, A. M. and Wagner, R., editors, *Database and Expert Systems Applications, DEXA, 10th International Workshop on Text-based Information Retrieval, TIR*, pages 49–53. IEEE.

Gazendam, L., Wartena, C., Malaisé, V., Schreiber, G., De Jong, A., and Brugman, H. (2009). Automatic annotation suggestions for audiovisual archives: Evaluation aspects. *Interdisciplinary Science Reviews, 34*, 2(3):172–188.

Girardot, J.-J. (2008). Evolution of the concept of territorial intelligence within the coordination action of the european network of territorial intelligence. *Ricerca e Sviluppo per le politiche sociali*, 1(1-2):11–29.

Girardot, J.-J. and Brunau, É. (2010). Territorial intelligence and innovation for the socio-ecological transition. In *9th. International conference of Territorial Intelligence, ENTI, Strasbourg*.

Grineva, M. P., Grinev, M. N., and Lizorkin, D. (2009). Extracting key terms from noisy and multitheme documents. In *Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009*, pages 661–670.

Herbaux, P. (2008). Tools for territorial intelligence and generic scientific methods. In *Internationa Annual Conference on Territorial Intelligence. Besançon: 16.16 October*.

Isaac, A. and Summers, E. (2009). Skos simple knowledge organization system primer. W3C Working Group Note. http://www.w3.org/TR/skos-primer/.

Jimenez, F., Fernández, I., and Menéndez, A. (2011). Los sistemas regionales de innovación: revisión conceptual e implicaciones en américa latina. In *Los Sistemas Regionales de Innovación en América Latina*. Banco Interamericano de Desarrollo, Washington.

Lundvall, B., editor (1992). *National Systems of Innovation: Towards a Theory of Innovation and Interactive Learning*. Pinter, London.

Lundvall, B.-A. (1998). Why study national systems and national styles of innovations? *Technology Analysis & Strategic Management*, 10(4):407 – 421.

Lundvall, B. A. and Christensen, J. L. (2003). Broadening the analysis of innovation systems–competition, organisational change and employment dynamics in the danish system. In P., C., Heitor, M., and BA, L., editors, *Innovation, Competence Building and Social Cohesion in Europe. Towards a Learning Society*, chapter Broadening the analysis of innovation systems–competition, organisational change and employment dynamics in the Danish system., pages 144–179. Cheltenham UK: Edward Elgar.

Lundvall, B.-A. and Johnson, B. (1994). The learning economy. *Journal of Industry Studies*, 1(2):23–42.

Malaisé, V., Gazendam, L., and Brugman, H. (2007a). Disambiguating automatic semantic annotation based on a thesaurus structure. In Hathout, N. and Muller, P., editors, *Actes de la 14e conférence sur le Traitement Automatique des Langues Naturelles (communications orales)*, pages 197–206, Toulouse. Association pour le Traitement Automatique des Langues.

Malaisé, V., Isaac, A., Gazendam, L., and Brugman, H. (2007b). Anchoring dutch cultural heritage thesauri to wordnet: two case studies. In *ACL 2007*, pages 57–63.

Medelyan, O. and Witten, I. H. (2005). Thesaurus-based index term extraction for agricultural documents. In *Proc. of the 6th Agricultural Ontology Service workshop*.

Mollo, M. (2009). The survey on territory research in europe,. In *International Conference of Territorial Intelligence, Papers on Tools and methods of Territorial Intelligence (MSHE)*, Besançon.

Nahapiet, J. and Ghoshal, S. (1998). Social capital, intellectual capital, and the organizational advantage. *The Academy of Management Review*, 23(2):242–266.

Nelson, R. R., editor (1993). *National Innovation Systems: A Comparative Study*. Oxford University Press, Oxford.

OECD and EUROSTAT (2005). *Oslo Manual: Guidelines for collecting and interpreting innovation data*. OECD Publising and European Commission. 3rd. edition.

Salavisa, I. and Vali, M. (2012). *Social Networks, Innovation and the Knowledge Economy*. Routledge.

Salton, G. and Buckley, C. (1987). Term weighting approaches in automatic text retrieval. Technical report, Cornell University.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.

Tiun, S., Abdullah, R., and Kong, T. E. (2001). Automatic topic identification using ontology hierarchy. In Gelbukh, A. F., editor, *Computational Linguistics and Intelligent Text Processing, Second International Conference, CICLing 2001, Mexico-City, Mexico, February 18-24, 2001, Proceedings*, volume 2004 of *Lecture Notes in Computer Science*, pages 444–453. Springer.

Turney, P. D. (2000). Learning algorithms for keyphrase extraction. *Inf. Retr.*, 2(4):303–336.

Wang, J., Liu, J., and Wang, C. (2007). Keyword extraction based on pagerank. *Advances in Knowledge Discovery and Data Mining*, 4426:857–864.

Wartena, C., Brussee, R., Gazendam, L., and Huijsen, W. (2007). Apolda: A practical tool for semantic annotation. In *Database and Expert Systems Applications, DEXA, 7th International Workshop on Text-based Information Retrieval, TIR*, pages 288–292. IEEE.

Weeds, J., Weir, D. J., and McCarthy, D. (2004). Characterising measures of lexical distributional similarity. In *COLING 2004, Proceedings of the 20th International Conference on Computational Linguistics*.