Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2014) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

CRIS 2014

Black Magic Meta Data - get a glimpse behind the scene

Thomas "Voldemort" Vestdam[a,]*, Henrik Steen "Saruman" Rasmussen[a], Marius "Sidious" Doornenbal[a]

[a]*Elsevier, Niels Jernes Vej 10, 9220 Aalborg East, Denmark*

Abstract

This paper presents how we utilise natural language processing techniques in order to "automagically" classify information stored in a CRIS, and aggregate the information in a researchers portfolio into a "fingerprint" describing a researchers research interest. Our approach exploits the fact that entities in a CRIS typically include some kind of text – most notable example being publication abstracts. We explain how the approach can result in automatic detailed classification of information, and argue how we can take advantage of such information in order to facilitate networking. Finally, we describe how we have realised the solution within our CRIS system.

Peer-review under responsibility of euroCRIS.

Keywords: CRIS systems; term extraction; auto-classification; fingerprinting; keywords; CERIF

1. Introduction

Even with a modern commercial CRIS system, researchers still have to do some manual and time-consuming work when it comes to inputting and maintaining meta-data about research (i.e. research information). Inputting meta-data into a CRIS system involves any activity needed to register information on publications, projects, activities, datasets, or any other relevant research. Maintenance of meta-data means adding additional information over time – information that was not available at the time the core data was registered in the system - e.g. uploading full-texts, re-classifying information, adding metrics (e.g. citations, impact factors, etc.), adding missing bibliographic information, or simply correcting erroneous information.

* *E-mail address:* T.Vestdam@Elsevier.com

Peer-review under responsibility of euroCRIS.

Surely, it would be desirable if your meta-data were *automagically* maintained and reasoned by, as if by use of black magic.

Unfortunately, as we all know, there is no such thing as real black magic - no ultimate automatisation - no matter what any computer scientist tells you. Luckily there are a number of small tricks that can be performed, and when combined, the result will be meta-data appearing in a CRIS, meta-data being maintained in the CRIS, and meta-data being reasoned about in the CRIS, as if where we using black magic.

In this paper we will give the reader insights into what happens behind the scenes in a modern commercial enterprise CRIS system that utilises natural language processing (NLP) techniques to alleviate researchers some of the tediousness of classifying their content, and helps to visualise content in a new way. First we motivate the usefulness of automatic classification functionality within CRIS, then we explain in more detail how we apply natural language processing and lift the veil on some of the principles behind our specific algorithm, and finally explain how the proposed functionality is actually implemented in our CRIS. We also propose a new structure in CERIF [8] that enables us to "classify" entities in CERIF using both keywords and classifications in a generic and structured way (compared to the current model).

2. The unfulfilled promise of the usefulness of classifications

We define a *classification* as a term within a controlled vocabulary, thesaurus or terminology, and we define a *keyword* as the more generic concept of an index term meant for information retrieval. Information retrieval can support many discoverability use cases, ranging form discoverability in the traditional sense via search engines, to more complex discoverability facilitating networking. Hence, a classification is considered a keyword in our context.

Facilitating networking is a particular hot topic, and among many things includes the desire to be able to find fellow researchers or research groups, that share common research interests, e.g. in order to seek collaborative funding or reviewers. One way to discover such "opportunities" could be to compare the research interests of the researchers. If these research interests are comprehensively represented as keywords, then you can use the keywords as input to a comparison algorithm. The algorithm can then assign a score that describe how similar the researcher's research interests are. However, it is difficult to describe a researcher's research interest solely by the use of keywords without a way to identify differences in importance between the individual keywords. A simple solution is to require that keywords are associated with a weight expressing the relevancy or importance of that given keyword relative to other keywords in the same list. The "only" problem is that it is not realistic to expect researchers to actively and manually maintain their research interests in form of comprehensive lists of weighted keywords within a CRIS.

Alternatively, if keywords were associated with the items in a researcher's *portfolio* – i.e. the researchers publications, projects, data sets etc., then the research interests could be deduced based on an aggregation of the individual items in the researchers portfolio (or parts of it). This would again require that the keywords on the items in the portfolio describe the actual contents of a given portfolio-item fully semantically and precisely.

However, as useful as keywords may be, it is still a huge task and burden to add and maintain keywords in a research information system. Hence, decorating and maintaining keywords on information in your CRIS is a manual task that laboriously must be performed by the researchers themselves. Some sources, like PubMed, Web of Science and Scopus, do supply keywords on publications, but that requires that these are actually used, and have always been used, when creating information in your CRIS. Even so, these sources might not supply keywords that are both consistent (e.g. over time) and comprehensive, and precise enough for our purpose. Furthermore, these sources only supply meta-data on publications – so, what about grants, projects, patents, activities, data sets, and even funding opportunities? Finally, the existing keywords may refer to different classification systems (controlled vocabularies) or worse, to no standardized system at all.

Our goal has been to extract enough information from text in order to be able to automatically classify that text – and, essentially classify what the text "is about" within the context of given a thesaurus (vocabulary). We are not dealing with an open interpretation of the text in order to classify it, but rather interested in answering questions like: "how is this text related to, say, MeSH terms?". Structured information of this type allows us to elevate this information to more aggregated information – for instance, about a set of publications that represents a researcher's

profile (or parts of it), and thereby automatically determine the research interest of that particular researcher. So in short, by applying some "black magic" we can automatically generate a list of classifications for any kind of entity (publications, projects, activities, data-sets, etc.) in our CRIS system that is decorated with text – for example in the form of a publication abstract, project description, data-set description, etc. Each classification is weighted to indicate its relevancy for a text. Furthermore, we generate a list of classification from each relevant classification system that we support (e.g. MeSH, Gesis, Humanities, Compendex, Geobase, Cambridge Math). Hence, all of this can be done on demand for all meta-data you have in your CRIS, including any legacy data from an old system you may have included in your CRIS.

3. Automatic fingerprinting using Natural Language Processing for concept extraction

The novelty in our approach to computer-assisted classification of content in a CRIS system is to apply Natural Language Processing (NLP) techniques to achieve this goal. Natural language processing [1], in our context, is the ability of a computer program to understand text as written by a human. More specifically, we are interested in computer-assisted understanding of text limited to being able to automatically assign semantics to a piece of text in the form of classifications, taken from a controlled vocabulary.

Natural language is the prototypical unstructured information - information that cannot be manipulated as such by computers. NLP tools perform the translation of unstructured information to a structured form. In deriving structured information from a text it does not suffice to list words or other selected strings occurring in the text: text has structure (syntax) and meaning. Words are by their very nature ambiguous: word meanings vary according to context and from speaker to speaker, and different words may express the same meaning.

3.1. The Elsevier Fingerprint Engine™

For the automated generation of semantic classifications on CRIS entities we employ the Elsevier Fingerprint Engine (EFE) – a concept annotation system that might be compared to several biomedical concept annotators as evaluated in [1]. There are other annotation systems – each specific to a scientific domain, but the requirement for a comprehensive CRIS system that covers all science domains, is that the fingerprint system should be applicable to all domains – ranging from engineering to medicine.

The concept annotation task performed by the EFE goes beyond simple term or entity recognition. The EFE uses various NLP techniques in order to deduce which key concepts a given piece of text is about – term annotation is just one of these techniques. The science of NLP techniques is a fast growing and well-researched field, and there are quite a few fairly mature tools that perform semantic analysis of texts.

The result of running our NLP algorithm on a piece of text is a fingerprint – the desired structured information. A fingerprint is set of pairs (concept, weight) – where concepts are terms from a given thesaurus, and the weight denotes the "confidence" in how well the concepts actually apply to the text that has been processed, or rather, how relevant that specific concept is for the input text. As noted, fingerprints can in turn be aggregated into fingerprints for a selection of publications, or even an entire researcher portfolio.

3.2. The fingerprinting process

The fingerprinting process is a concept annotation process that can be applied to texts of any size – from single lines to abstracts, or, conceivably, to full text articles. The process involves the consecutive execution of a number of NLP steps, where each step builds on the results of previous steps. A modular design of the EFE, comparable to that of similar frameworks as UIMA [2] or GATE [3] was chosen to allow for tailor-cut text-processing pipelines to meet the specific requirements of each situation and science domain.

It obviously carries too far in this context to detail the steps for text analysis for any specific domain. However, for all domains the processing pipeline includes steps of tokenization, input analysis and normalization, expansion of abbreviations and coordinations (similar to [4] and [5]), a number of entity annotations (names, institutions, and citations) and part-of-speech tagging. These steps are preparatory relative to the term annotation step, in which the text is scanned for the occurrence of terms as defined in the target thesaurus or vocabulary that is designated as the

relevant set of concepts for the domain. During the term annotation step, textual variations such as normalization and spelling differences, punctuation and word order variations, and part-of-speech tags are either ignored or taken into account –depending on the nature of the terms sought for. For instance, while word order is very relevant for the term pair *Host vs. graft reaction* and *Graft vs. host reaction*, stringent ordering conditions are relaxed for most other terms. After the term annotation task has been performed, annotated terms are evaluated in a number of disambiguation steps, which establish certainty on the question whether an annotated term candidate really designate the concept as it is defined in the target vocabulary. Word sense disambiguation (WSD) is an indispensable component of NLP solutions that claim to provide semantic annotation. For an overview of the field, see [6] and [7]. The EFE employs disambiguation techniques based on pattern-based rules, as well as statistical co-occurrence methods, unification methods and thesaurus-based co-occurrence methods.

The Fingerprint Engine maps texts onto a domain vocabulary. The power of using thesauri is that they offer a structured view of the conceptual space for a scientific domain, not only offering possibly multiple synonyms for single concepts, but also definitions for those concepts and relations between concepts. The downside of using thesauri is that it is not trivial to get a single view on multidisciplinary research output. This problem is not only caused by the mundane problem that there are no comprehensive high-quality unifying vocabularies that are commonly accepted, but also fundamental in nature – there is no way getting around the fact that *morphology* simply means something different to a medical researcher and a linguist. In any case, if we are to present a unified profile, the domain source of concepts must be annotated to make the profile truly semantic.

The most important basis for profiling researchers output is formed by the abstracts of published articles. Abstracts of papers are typically a summary of the entire paper of controlled length, resulting in information-rich fingerprints of comparable lengths. It is natural to think that analysing the full text yields more information – however, the different parts of a publication may be of a very different nature: for instance, the *methods*, *future work* or *related work* sections of a publication will yield terms that are not relevant to the essence of the publication. Hence, if a machine is to deduct if two publications are on similar topics, it might conclude that they are similar, solely on the grounds that their methodology is very similar. The Fingerprint Engine is currently also applied to funding awards, funding applications and funding opportunities.

3.3. Future challenges

As noted the current weakness of the EFE is its inability to present multidisciplinary profiles in a single unified view. This is not necessarily a weakness of the Fingerprint Engine itself; also, this weakness has the advantage that the view of a researcher's output is associated with the domain of his or her activities – should a researcher have a profile in both engineering and life sciences, he or she may be active in bioengineering and is discoverable by linking to other entities from both domains. We are actively seeking to overcome the limitations in this respect in two directions. The first solution to this problem is to (still) use controlled vocabulary profiles as a presentation means – to retain the 'pretty view' on a researcher's output – while using a noun phrase or other feature set index in the background for discovery and linkage purposes. At the same time, we pursue a strategy where we set apart the truly domain specific concepts from the concepts that are constant across domains: entities such as geographical locations, species names and chemicals.

4. Fingerprints in action

We have implemented support in our CRIS for storing fingerprints, displaying fingerprints as well as reasoning about fingerprinted information. The main focus has been on publications and researcher profiles (see figure 1). All publication abstracts are fingerprinted and researchers are automatically given a fingerprint based on the aggregation of fingerprinted items in their portfolio.

We have also implemented search functionality that allows visitors on an institutions research web site to search for researchers working on specific research areas (with a given thesaurus). This search can be performed among the researchers at a given institution, or among all institutions that participate in a community of experts.

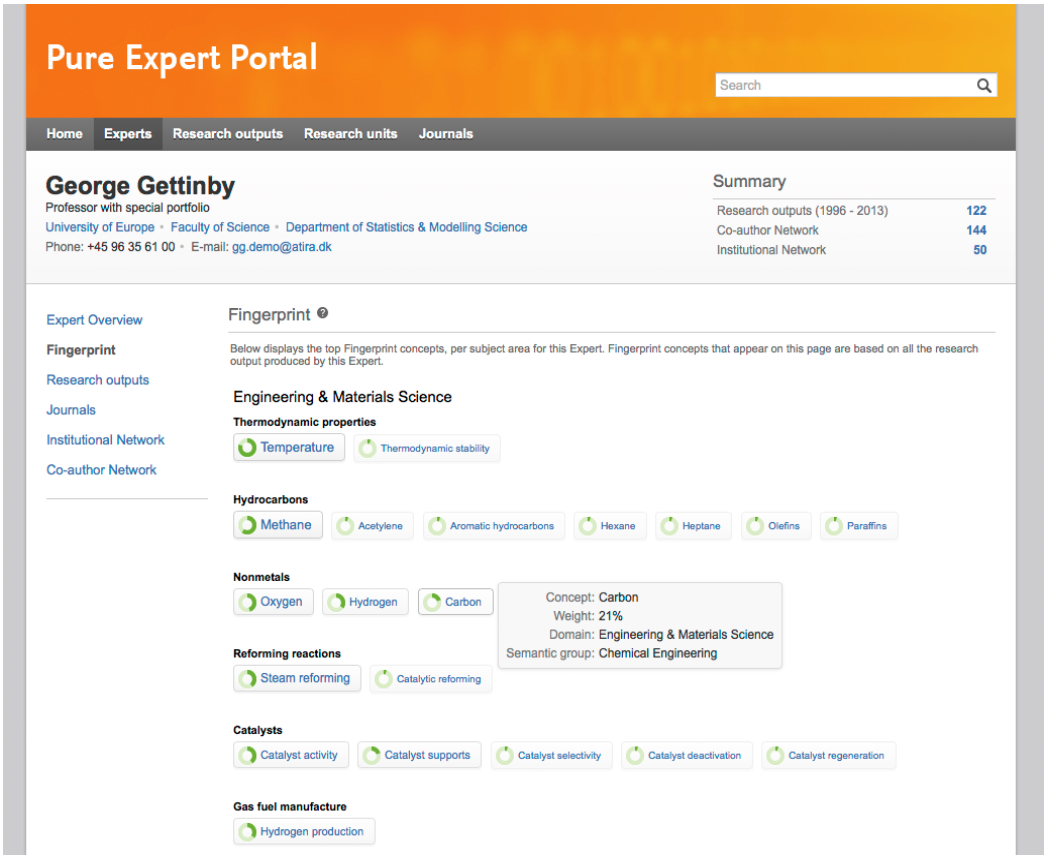

Fig. 1. Fingerprints in action on a researcher's public profile

Automatic generation of fingerprints is not perfect – for example if a researcher's portfolio contain publications within research areas that the researcher is no longer interested in. Researchers therefore have options for modifying their fingerprint(s) - either by removing or adding weighted terms.

5. Fingerprints and generic keywords in CERIF

Representing fingerprints in CERIF [8] is unfortunately not straightforward. A fingerprint is essentially a triple (thesaurus, thesaurus version, concepts), where concepts are a set of pairs on the form (term, weight). In CERIF lingo that is (classification scheme, "version") and (classification, fraction). There does not exist a construct in CERIF that allows you to represent this kind of information. Obviously this is only relevant if the idea of fingerprints is deemed of general purpose, but there are good reasons to take a step back, and have a look for a more generic structure for keywords in CERIF that can provide a bit more structure around keywords, and this construct would also be able to represent fingerprints.

Classifications in CERIF are a general concept that is utilised for applying all thinkable kinds of semantics to entities in CERIF. So, classification of a publication in the context of MeSH terms would be represented as cfResPubl_Class entities, alongside with any other relevant classifications of the publication, such as publication type, review status, publication status etc. The different classifications are of cause distinguishable based on the reference in cfResPubl_Class to a specific classification schemes.

Uncontrolled keywords are represented in cfResPublKeyw as a comma separated string of individual keywords within a given language (and translation). However, keywords are often relevant in the context of a controlled vocabulary, as a means to supply more precision than what a given term in the vocabulary allows.

A re-design of how keywords are represented in CERIF could yield a more generic solution, and at the same time provide more structure by encapsulating "keywords" in general. The following is deliberately not in "CERIF syntax", as this is not an implementation proposal, but a high-level description of what we are proposing should be implemented in CERIF.

We suggest introducing thee new "entities" that each represents the concept of a *Free Text Keyword*, a *Controlled Keyword* and a *Keyword Group.*

The attributes of a *Free Text Keyword* would be a keyword (text), fraction and an indication of language and translation – i.e. similar to any other multiple language entities in CERIF.

The attributes of a *Controlled Keyword* would be a classification scheme, classification, fraction and a list of free text keywords – i.e. similar to any classifying link entity in CERIF, except that it can be linked with a number of Free Text Keywords.

Finally the attributes of the *Keyword Group* would be a classification scheme and a list of *Controlled Keywords*.

Hence, if the concepts outlined above are translated to formal CERIF, then CERIF would have a generic model for grouping keywords in a container that can contain "keywords" from both controlled and uncontrolled vocabularies, and a concept that allows a classification from a controlled vocabulary to be further specified – and, "fingerprints" could be represented as a Keyword Group.

6. Conclusion

We have demonstrated how we have applied Natural Language Processing techniques to automatically deduce sematic information based on text stored in our CRIS system. This allows us to generate fingerprints that - in the context a given thesaurus - semantically classify a given item stored in the CRIS. Fingerprints contain both classification terms, and a weight that denotes the "confidence" in how well concepts actually apply to the text that has been processed. Finally, we can aggregate the fingerprints within a researchers portfolio in order to provide a fingerprint for the researcher that describes his or hers specific research interests.

If the techniques presented in this paper are applied in a CRIS then it is indeed possible to automatically classify and re-classify information over time - hence facilitating networking, search scenarios, comparison and similarity analysis, etc. However, the approach does require some text to be available in your CRIS – such as publication abstracts or descriptive text usually available for funding awards, funding applications, funding opportunities and projects. Text, that is possible to find and store in a CRIS (even automatic), and many institutions already have such information in their CIRS.

Our specific NLP algorithm is proprietary, and is efficient and precise – yet, the ideas presented in this paper can be applied by anyone who has access to NLP tools, or by utilising openly available tools.

We have also proposed an extension of CERIF that allows concepts like fingerprints and generic keywords to be represented in CERIF in a structured way (as opposed to the current model).

References

1. Funk C., et al. Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters BMC Bioinformatics 2014, 15:59, doi:10.1186/1471-2105-15-59
2. http://uima.apache.org/
3. Cunningham H., Maynard D., Bontcheva K. and Tablan V. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In: proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics, 2002
4. Kang N., Singh B., Afzal Z., van Mulligen EM. and Kors JA. Using rule-based natural language processing to improve disease normalization in biomedical text. *Journal of the American Medical Informatics Association*, 2012
5. Ao, H. and Takagi, T. ALICE: an algorithm to extract abbreviations from MEDLINE. *Journal of the American Medical Informatics Association*, 12, 2005, 576-586
6. Navigli R. Word sense disambiguation: ACM Computing Surveys, 41(2), ACM Press, 2009, 1-69.
7. Navigli R. A Quick Tour of Word Sense Disambiguation, Induction and Related Approaches. In: Proceedings of the 38th Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM), 2012, 115-129, Spindleruv Mlyn, Czech Republic
8. The Common European Research Information Format (CERIF); a EU Recommendation to Member States: http://cordis.europa.eu/cerif/, http://www.euroCRIS.org/