Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 33 (2014) 239 – 244

Procedia
Computer Science

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.

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

doi:10.1016/j.procs.2014.06.038

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