Contextual Semantic: A Context-aware Approach for Semantic Web Based Data Extraction from Scientific Articles

Deniss Kumlander
Department of Informatics, TTU
Raja 15, 12618
Tallinn, Estonia
e-mail: kumlander@gmail.com

Abstract—**The paper explores whether the semantic context is good enough to cope with ever increasing number of available resources in different repositories including the web. Here a problem of identifying authors of scientific papers is used as an example. A set of problem still do arise in case we apply exclusively the semantic context. Fortunately contextual semantic can be used to derive more information required to separate ambiguous cases. Semantic tags, well-structured documents and available databases of articles do provide a possibility to be more context-aware. Under the context we use co-authors names, references and headers to extract key-words and identify the subject. The real complexity of the considering problem comes from the dynamical behaviour of authors as they can change the topic of the research in the next paper. As the final judge the paper proposes applying words usage patterns analysis. Final the contextual intelligence engine is described.**

Keywords-Semantic net, library analysis, context-aware.

I. INTRODUCTION

The quick evolution of technologies, globalisation of the world and quick growth of the educational level in countries that are just now merging the world scientific society generated the huge number of information and articles. The growth of those papers models the growth of published document in the web and certainly exceeds expectations of researchers. In this situation analysis of those works and consequently the search of key papers by keywords become more and more complicated and nowadays nearly impossible.

The first task that librarians, including online libraries and search engines, have to solve is the correct linking of articles to authors. This task is trivial on small amount of articles in limited training amount of those become fairly complex on real examples. The paper do stop on this in more details in the forth chapter. In order to solve this problem the semantic web, described in the chapter two, was mainly used so far. Unfortunately this important technology is not good enough to cope with all kind of complexities, but still do provide an excellent basis for further processing of papers. Here, we do propose to strength the semantic net approach with context-aware systems approaches. Saying that we do warn the reader that the semantic context and semantic and context should be distinguished as those do represent completely different approaches as we will demonstrate in more details in the third chapter.

The firth chapter does specify the proposed approaches including problems and solutions and the final chapter concludes the paper.

II. SEMANTIC WEB

The Semantic web [1] idea is the further evolution of the web proposed in 1990th where the global network doesn't only stores the data, but also contains semantic information and relates different documents on this basis. Documents in the web are published using HTML, which does provide certain information in some extend, but is first of all limited and secondly is a mixture of design, styling, types and other kind of tags requiring to present the document correctly to web browsers, but not really good enough to process any document efficiently. The semantic web, basing on technologies like RDL, OWL and XML allow giving additional information to data (text of the document) described in the published document and understand at least key information like for example the name of the author, the main topic of the document and so forth

The semantic net can be seen as a further generalisation of the semantic web, as documents can be stored in any kind of repository, which is not always the Web.

The semantic net shifts abilities to process information stored in documents to the new level. If the old data representation was primary used for machines to process and display information for humans, then in the semantic net the information can be processed, displayed and understood by the machine. Basing on this understanding the computer can do additional activities, for example fetching additional links which are similar and are likely to be interested for the human requested the current document.

III. CONTEXT AWARE SYSTEMS

A context aware software development principal defines the computer reaction can be organized using more than just direct data input. It could and probably should consider circumstances under which the input is defined, sense the environment in

T. Sobh, K. Elleithy (eds.), *Innovations in Computing Sciences and Software Engineering*,
DOI 10.1007/978-90-481-9112-3_40, © Springer Science+Business Media B.V. 2010

which the request is formulated and so forth. In the result the system response will be different for the same request depending on those extra (request context) parameters.

Some authors [5, 6] say that this behaviour can be described by the system ability to adapt to specific contexts under which the system operates or reacts on the request and ability to collect and recognize the context information.

Consider for example web applications, which do normally react on the request, but could collect and modify behaviour depending on the following extra parameters:

- Previous requests that are not considered as direct input to the current one. For example so far collected preferences of a user without asking him to define those directly like previously made searches in case of search engines etc.
- Internet browser type. A request made using IE browser should be responded differently in compare to Firefox as those browsers standards (JavaScript, HTML) etc. are different so the response will be rendered differently;
- Accessing device – obviously the page should look different in case the user accesses the page using mobile version of the device due space, keyboard etc restrictions;
- Vary the response depending on geo location of the requester (country, time-zone, language);
- Connection speed – in order to provide either a short and restricted answer or expanded and large one;

… and so forth.

The example above deals with hardware context used in most cases, but obviously it is not the only context that can be acquired. Following the same principals, it is possible to use history of communication or different other messages as a context of the current communication.

The context-aware approach is used to increase relevancy of the response (service) provided by the system and target to bridge the gap between formal and [7] human-human type interaction. Generally context-aware application contain 3 types of sub-activity elements: they should detect context, interpret it and react on it, although the first activity is sometimes delegated to other systems (alternatively it can be said that the system relies on other one to collect context and provide it and so depends on the front-system in terms of how well it can adapt as the smaller context is the less flexibility the system could provide).

IV. Sample Case

The modern society provides much more opportunities for researches and consequently for publishing novel approaches in forms of articles. The overall speed of knowledge transfer is permanently increasing allowing collaborating and picking up the current knowledge quickly and efficiently. It does provide much better starting position for any research to be done than ever before. Fifty years ago the post looked like an incredible invention in compare to what scientists had in the 19th century been isolated into local country communities. Only major invention was floating more or less freely having still a sufficient shift in time between been acquired and transferred to any other community. The same shift has been done once again with invention of Internet system 25 years ago allowing much faster access information and what is much more important much easier publish information. Nowadays search engines do simplify the process of searching information even more grouping it, tagging by key words, so the user doesn’t have to know the exact address to obtain information from, but could use keywords or paper names in the search process.

The growth of population and involvement into active science other countries used to be called the „third world” sufficiently increased the number of contributions made each year.

Imagine that our task is to structure articles by authors. At the first glance the process is obvious. Current articles’ format requirements contain semantic tag to be used to identify authors of the paper. Therefore in most cases authors can be easily extracted as well as other information like references, keywords, topics of the papers orienting on headers of different level.

The real problem comes with the overall world involvement and the size of available papers that makes exactly similar names of totally different persons to be quite common nowadays. In the next chapter we will try to resolve this problem applying context-aware systems techniques. We will call such persons below in the paper – “co-persons”.

V. Semantic Contextual Web

Is semantic enough to derive the correct result? The first answer on this question has been got already in early days of artificial intelligence algorithms implementation when automatic translation assistants were developed. Clearly semantic of any statement can be interpreted differently and so the meaning is very ambiguous. The context is what defines the meaning of most concrete statements. Following the best practises from this (translation) area the first level of deriving context will be exploring the neighbourhood of the statement by examining other statements. During this process the algorithm searches for interdependencies and attempts to narrow down the topic. This process can be called a semantic context.

Unfortunately the semantic context is not enough to derive required information in many cases. For example if we do analyse messages transferred during communication between two persons then backgrounds of communicating parties are required parts of the puzzle, which is missing as in not directly send (and so written down) during the communication messages exchange. Therefore we need to consider other context of the information, which is not directly linked to the semantic of language or parsed tags. In other words we propose also using contextual semantic.

Returning to our task defined earlier, we would like to reformulate it slightly by providing the background. The problem is not about processing one and only one article. On