Interlinking e-Learning Resources and the Web of Data for Improving Student Experience

Antonella Carbonaro

Department of Computer Science, University of Bologna, Italy
antonella.carbonaro@unibo.it

The Web provides not only several data sources with useful and relevant information with e-Learning purposes, but also information that is not easy to retrieve. The web of linked data is a repository based on semantic technologies. Several researchers have been oriented to this kind of interoperable e-Learning repositories and establish that the Linked Data approach has the potential to fulfill the e-Learning vision of Web-scale interoperability of e-Learning resources as well as highly personalized and adaptive e-Learning applications. The paper presents an automatic concept extraction system used to improve personalized searching framework. It could be considered as one possible instance of a more general concept concerning the transition from the Document Web to the Document/Data

for citations:
Carbonaro A. (2012), *Interlinking e-Learning Resources and the Web of Data for Improving Student Experience,* Journal of e-Learning and Knowledge Society, v.8, n.2, 33-44. ISSN: 1826-6223, e-ISSN:1971-8829

Web and the consequent managing of these immense volumes of data.

1 Introduction

The collection and the use of data on the Internet with e-learning purposes are tasks made by many people every day, because of their role as teachers or students. The Web provides several data sources with relevant information that could be used in educational frameworks, but the information is widely distributed, or poorly structured. Moreover, to have a effectiveness personalized experience standard keyword search has a very limited effectiveness; for example, it cannot filter for the type of information, the level of information or the quality of information. These situations involve a difficult search of e-learning resources, and therefore a lot of time invested, because the search process is completely executed by humans, even with some tasks (reasoning, selecting, using resources, bookmarking, and so on) could be executed by computers.

The Semantic Web offers a generic infrastructure for interchange, integration and creative reuse of structured data, which can help to cross some of the boundaries that Web 2.0 is facing. Currently, Web 2.0 offers poor query possibilities apart from searching by keywords or tags. There has been a great deal of interest in the development of semantic-based systems to facilitate knowledge representation and extraction and content integration (Henze *et al.*, 2009; Bighini *et al.*, 2004). Semantic-based approach to retrieving relevant material can be useful to address issues like trying to determine the type or the quality of the information suggested from a personalized environment. Potentially, one of the biggest application areas of content-based exploration might be personalized searching framework (e.g., Pickens; Freyne & Smyth, 2004). Whereas search engines provide nowadays largely anonymous information, new framework might highlight or recommend web pages related to key concepts. We can consider semantic information representation as an important step towards a wide efficient manipulation and retrieval of information (Calic, 2005; Carbonaro, 2006; Bloehdorn *et al.*, 2004). In the digital library community a flat list of attribute/value pairs is often assumed to be available. In the Semantic Web community, annotations are often assumed to be an instance of an ontology. Through the ontologies the system will express key entities and relationships describing resources in a formal machine-processable representation. An ontology-based knowledge representation could be used for content analysis and object recognition, for reasoning processes and for enabling user-friendly and intelligent multimedia content search and retrieval.

Although the Semantic Web is a Web of Data, it is intended primarily for humans; it would use machine processing and databases to take away some of the burdens we currently face so that we can concentrate on the more important

things that we can use the Web for.

The idea behind Linked Data (Bizer *et al.*, 2009) is using the Web to allow exposing, connecting and sharing linking data through dereferenceable URIs on the Web. The goal is to extend the Web by publishing various open datasets as RDF triples and by setting RDF links between data items from several data sources. Using URIs, everything can be referred to and looked up both by people and by software agents. Berners-Lee expose the basis of Linked Data techniques and highlight the differences between the two modes of web information: the web of hypertext, and the web of data. Both are constructed with documents on the web, but the web of data is simply about using the Web to create typed links between data from different sources; technically, Linked Data refers to data published on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external data sets, and can in turn be linked to from external data sets. Therefore, this information could be consumed by people anytime as resources with e-learning purposes.

The paper is organized as follows: firstly, we introduces personalized searching framework as one of the possible application areas of automatic concept extraction systems. Then, we describe the summarization process, providing details on system architecture, used methodology and tools. Subsequently, we introduce some different available open metadata standards. Finally, we provide some considerations on case study and experimental results.

2 Personalized Searching Experience

In personalized searching frameworks, standard keyword search is of very limited effectiveness. For example, it does not allow users and the system to search, handle or read concepts of interest, and it doesn't consider synonymy and hyponymy that could reveal hidden similarities potentially leading to better retrieval. The advantages of a concept-based document and user representations can be summarized as follows: (i) ambiguous terms inside a resource are disambiguated, allowing their correct interpretation and, consequently, a better precision in the user model construction (e.g., if a user is interested in computer science resources, a document containing the word 'bank' as it is meant in the financial context could not be relevant); (ii) synonymous words belonging to the same meaning can contribute to the resource model definition (for example, both 'mouse' and 'display' brings evidences for computer science documents, improving the coverage of the document retrieval); (iii) synonymous words belonging to the same meaning can contribute to the user model matching, which is required in recommendation process (for example, if two users have the same interests, but these are expressed using different terms, they will be considered overlapping); (iv) finally, classification, recommendation and

sharing phases take advantage of the word senses in order to classify, retrieve and suggest documents with high semantic relevance with respect to the user and resource models.

Because information is costly (in terms of time) to download, displays of result lists should be optimized to make the process of browsing more effective. Using implemented tools, searchers found relevant documents more efficiently and effectively and they found relevant documents that otherwise went undiscovered.

A way to support the user in having a suitable search experience is based on the use of a concept extraction technique that tends to increase the precision of retrieval, and thus is key in supporting focused search; moreover, the use of concepts based on some notion of similarity supports search by increasing the recall of retrieval, by suggesting possibly relevant items by utilizing proposed techniques.

3 Concept extraction

Text summarization has been an interesting and active research area since the 60's. The definition and assumption are that a small portion or several keywords (concepts) of the original long document can represent the whole informatively and/or indicatively. Reading or processing this shorter version of the document would save time and other resources (White *et al.*, 2009). This property is especially true and urgently needed at present due to the vast availability of information. Concept-based approach to represent dynamic and unstructured information can be useful to address issues like trying to determine the key concepts and to summarize the information exchanged within a personalized environment.

Researchers have reported a rich collection of approaches for automatic document summarization to enhance those provided manually by readers or authors as a result of intellectual interpretation. One approach is to provide summary creation based on a natural language generation (as investigated for instance in the DUC and TREC conferences); a different one is based on a sentence selection from the text to be summarized, but the most simple process is to select a reasonable short list of words among the most frequent and/or the most characteristic words from those found in the text to be summarized. So, rather than a coherent text the summary is a simple set of items.

From a technical point of view, the different approaches available in the literature can be considered as follows. The first is a class of approaches that deals with the problem of document classification from a theoretical point of view, making no assumption on the application of these approaches. These include statistical (McKeown *et al.*, 2001), analytical (Brunn *et al.*, 2001),

information retrieval (Aho *et al.*, 1997) and information fusion (Barzilay *et al.*, 1999) approaches. The second class deals with techniques that are focused on specific applications, such as baseball program summaries (Yong Rui *et al.*, 2000), clinical data visualization (Shahar & Cheng, 1998) and web browsing on handheld devices (Rahman *et al.*, 2001; NIST) reports a comprehensive review.

4 Summarization Process

Potentially, one of the biggest application areas of content-based exploration might be personalized searching framework (e.g., Bighini *et al.*, 2004; Pickens *et al.*). Whereas today's search engines provide largely anonymous information, new framework might highlight or recommend web pages or content related to key concepts. We can consider semantic information representation as an important step towards a wide efficient manipulation and discovery of information (Freyne & Smyth, 2004; Calic *et al.*, 2005; Carbonaro, 2006). In the digital library community a flat list of attribute/value pairs is often assumed to be available. In the Semantic Web community, annotations are often assumed to be an instance of an ontology. Through the ontologies the system will express key entities and relationships describing resources in a formal machine-processable representation. An ontology-based knowledge representation could be used for content analysis and object recognition, for reasoning processes and for enabling user-friendly and intelligent multimedia content exploration and retrieval.

Therefore, the semantic Web vision can potentially benefit from Information Retrieval, Information Extraction, Content Analysis and Lexicography applications, as it inherently needs domain-oriented and unrestricted sense disambiguation to deal with the semantics of documents, and enable interoperability between systems, ontologies, and users.

The approach presented in this chapter produce a set of items, but involves improvements over the simple set of words process in two means. Actually, we go beyond the level of keywords providing conceptual descriptions from concepts identified and extracted from the text. We propose a practical approach for extracting the most relevant keywords from the forum threads to form a summary without assumption on the application domain and to subsequently find out concepts from the keyword extraction based on statistics and synsets extraction. Then semantic similarity analysis is conducted between keywords to produce a set of semantic relevant concepts summarizing actual forum significance.

In this context, a concept is represented with a Wikipedia article. With

millions of articles and thousands of contributors, this online repository of knowledge is the largest and fastest growing encyclopedia in existence. The problem described above can then be divided into three steps:

- Mapping of a series of terms with the most appropriate Wikipedia article (disambiguation).
- Assigning a score for each item identified on the basis of its importance in the given context.
- Extraction of n items with the highest score.

In order to substitute keywords with univocal concepts we have to build a process called Word Sense Disambiguation (WSD). Given a sentence, a WSD process identifies the syntactical categories of words and interacts with an ontology both to retrieve the exact concept definition and to adopt some techniques for semantic similarity evaluation among words. We use MorphAdorner (Burns & Philip) that provides facilities for tokenizing text and WordNet (Fellbaum), one of the most used ontology in the Word Sense Disambiguation task.

The methodology used in this application is knowledge-based, it uses Wikipedia as a base of information with its extensive network of cross-references, portals, categories and info-boxes providing a huge amount of explicitly defined semantics.

To extract and access useful information from Wikipedia in a scalable and timely manner we use the Wikipedia Miner toolkit [http://wikipedia-miner.sourceforge.net/] including scripts for processing Wikipedia dumps and extracting summaries such as the link graph and category hierarchy.

In this chapter we focus on DBpedia (Bizer *et al.*, 2009), that is one of the main clouds of the Linked Data graph. DBpedia extracts structured content from Wikipedia and makes this information available on the Web; it uses the RDF to represent the extracted information. It is possible to query relationships and properties associated with Wikipedia resources (through its SPARQL endpoint), and link other data sets on the web to DBpedia data.

The whole knowledge base consists of over one billion triples. DBpedia labels and abstracts of resources are stored in more than 95 different languages. The graph is highly connected to other RDF dataset of the Linked Data cloud. Each resource in DBpedia is referred by its own URI, allowing to precisely get a resource with no ambiguity. The DBpedia knowledge base is served as Linked Data on the Web. Actually, various data providers have started to set RDF links from their data sets to DBpedia, making DBpedia one of the central interlinking-hubs of the emerging Web of Data.

Compared to other ontological hierarchies and taxonomies, DBpedia has the advantage that each term or resource is enhanced with a rich description including a textual abstract. Another advantage is that DBpedia automatically

evolves as Wikipedia changes. Hence, problems such as domain coverage, content freshness, machine-understandability can be addressed more easily when considering DBpedia. Moreover, it covers different areas of the human knowledge (geographic information, people, films, music, books, …); it represents real community agreement and it is truly multilingual.

5 e-Learning Metadata Standards and Practices

E-learning standards provide support especially to process educational resources into an interoperable manner. Some standards provide metadata specification for describing the properties of LOs (ARIADNE, DCMI, IEEE-LOM, ADL), others for describing the structure on content (AICC). As well, standards like IMS and ADL/SCORM (Sharable Content Object Reference Model) handle both metadata specification and content structure modeling (Milne & Witten, 2009).

In order to acquire interoperability with respect to the semantic description of educational resources, some semantic metadata should be additionally defined, by using standards specific to the Semantic Web. Various standards were defined, focused on specific information type description (DCMI, RSS, Atom, FOAF, DOAP, …). It is also possible to embed semantic metadata into Web resources to convey the meaning of the document itself, instead of collecting them into separated documents through microformats or through RDFa, which provide support in addition for metadata interlinking.

However, the combination between e-learning standards and Semantic Web standards is a difficult issue. For example, although a vast amount of educational content and data is shared on the Web in an open way, the integration process is still costly as different learning repositories are isolated from each other and based on different implementation standards (de Santiago & Raabe, 2010).

To overcome these problems, some educational institutions started to expose their data based on Linked Data principles, however these efforts mainly focus on exposing individual data while interlinking with 3rd party data is not yet within the primary scope.

Considerations

The work described in this chapter represents some initial steps in exploring automatic concept extraction in semantic summarization process. It could be considered as one possible instance of a more general concept concerning the transition from the Document Web to the Document/Data Web and the consequent managing of these immense volumes of data. The community of linked

data provides data sets that are already connected, and this information could be consumed by people anytime as resources with e-learning purposes.

Indeed, advances in search need to do more than simply improve the syntactic keyword matching process and can be used, for example, in new search scenarios, including when the users are (a) unfamiliar with a domain and its terminology, (b) unfamiliar with a system and it's capabilities, or (c) unfamiliar with the full detail of their task or goal.

Summarization can be evaluated using intrinsic or extrinsic measures; while the first one methods attempt to measure summary quality using human evaluation, extrinsic methods measure the same through a task-based performance measure such the information retrieval-oriented task. In our experiments we utilized intrinsic approach analyzing (Jones, 2007) as document and (Milne & Witten, 2008; Suchanek *et al.*, 2008; Yu *et al.*, 2007; Kittur *et al.*, 2009; Zesch *et al.*, 2008; Wolf & Gurevych, 2010; Milne & Witten, *op. cit.*; Schonhofen, 2009; Medelyan *et al.*, 2008; Amiri *et al.*, 2008; Mihalcea & Csomai, 2007) as corpus.

This experiment is to evaluate the usefulness of concept extraction in summarization process, by manually reading whole document content and comparing with automatic extracted concepts. The results show that automatic concept-based summarization produces useful support to information extraction. The extracted concepts represent a good summarization of document contents.

For example, we evaluated the influence of chosen window size using 605 terms to be disambiguated. The results are showed in Table 1.

TABLE 1
Change in precision and recall as a function of window size

Window	Copertura (C)	Precisione (P)	C%	P%
4	438	268	72.39	61.18
8	461	357	76.19	77.44
12	470	355	77.68	75.53
16	475	369	78.51	77.68
20	480	362	79.33	75.41

Using the best choice of parameter values we obtain the following percentages in precision and recall.

TABLE 2
Change in precision and recall using the showed set of parameter values

window	minScore	minRelatednessToSplit
8	0.18	0.2

Copertura (C)	Precisione (P)	C%	P%
444	358	73.38	80.63

Finally, given the document [MW08], Table 3 shows the ten most representative articles automatically extracted from the system.

While the initial results are encouraging, much remains to be explored. For example, many disambiguation strategies with specific advantages are available, so designers now have the possibility of deciding which new features to include in order to support them, but it is particularly difficult to distinguish the benefits of each advance that have often been shown independent of others.

TABLE 3
Automatically extracted articles representing [MW08]

Listing 6.16: I primi dieci articoli che rappresentano il documento [MW08].

```
1 Language
2 System
3 Category theory
4 Knowledge
5 Word
6 Datum (geodesy)
7 Concept
8 WordNet
9 Application software
10 Accuracy and precision
```

It would also be interesting to apply the showed method using a different knowledge base, for example YAGO (but always derived from Wikipedia) and use a different measure of relationship between concepts considering not only the links belonging to articles but also the entire link network. That is, considering Wikipedia as a graph of interconnected concepts, we could exploit more than one or two links.

While the initial results are encouraging, much remains to be explored. For example, many search strategies with specific advantages are available, so designers now have the possibility of deciding which new features to include in order to support them, but it is particularly difficult to distinguish the benefits of each advance that have often been shown independent of others.

Taking into account the amount of data located on the internet and the opportunity to use the datasets currently connected in the linked data community and to make connections between resources that provide usable information for e-Learning purposes, we could make the web a more interesting place, and also a relevant tool for e-Learners, in order to improve their experience in searching e-Learning resources.

REFERENCES

Henze N., Dolog P., Nejdl W. (2004), *Reasoning and Ontologies for Personalized E-Learning in the Semantic Web*, Educational Technology & Society, 7 (4), 82-97.

Bighini C., Carbonaro A. (2004), *InLinx: Intelligent Agents for Personalized Classification, Sharing and Recommendation*, International Journal of Computational Intelligence. International Computational Intelligence Society. 2 (1).

Pickens J., Golovchinsky G., Shah C., Qvarfordt P., Back, M., *Algorithmic Mediation for Collaborative Exploratory Search*. To appear in Proceedings of SIGIR

Freyne J., Smyth B. (2004), *Collaborative Search: Deployment Experiences*, in The 24th SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence. Cambridge, UK, pp. 121-134.

Calic J., Campbell N., Dasiopoulou S., Kompatsiaris Y. (2005), *A Survey on Multimodal Video Representation for Semantic Retrieval*, in the Third International Conference on Computer as a tool, IEEE.

Carbonaro A. (2006), *Defining Personalized Learning Views of Relevant Learning Objects in a Collaborative Bookmark Management System*, In Z. Ma (Ed.), Web-based Intelligent ELearning Systems: Technologies and Applications (pp. 139-155). Hershey, PA: Information Science Publishing.

Bloehdorn S., Petridis K., Simou N., Tzouvaras V., AvrithisY., Handschuh S., Kompatsiaris Y., Staab S., Strintzis M. G. (2004), *Knowledge Representation for Semantic Multimedia Content Analysis and Reasoning*, in Proceedings of the European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology.

Bizer C., Heath T., Berners-Lee T. (2009), *Linked data - the story so far*. International Journal on Semantic Web and Information Systems, 5(3):1.

White R. W., Roth R. (2009), *Exploratory Search: Beyond the Query-Response Paradigm*. Morgan & Claypool.

McKeown K., Barzilay R., Evans D., Hatzivassiloglou V., Kan M., Schiffman B., Teufel S. (2001), *Columbia Multi- Document Summarization: Approach and Evaluation*. Workshop on Text Summarization.

Brunn M., Chali Y., Pinchak C. (2001), *Text Summarization Using Lexical Chains*. Work. on Text Summarization.

Aho A., Chang S., McKeown K., Radev D., Smith J., Zaman K. (1997), *Columbia Digital News Project: An Environment for Briefing and Search over Multimedia*. Information J. Int. J. on Digital Libraries, 1(4):377-385.

Barzilay R., McKeown K. Elhadad M. (1999), *Information fusion in the context of multi-document summarization*. In Proc. of ACL'99.

Yong Rui Y., Gupta A., Acero, A. (2000), *Automatically extracting highlights for TV Baseball programs*. ACM Multimedia, Pages 105-115.

Shahar Y., Cheng C. (1998), *Knowledge-based Visualization of Time Oriented Clinical Data*. Proc AMIA Annual Fall Symp., pages 155-9.

Rahman A., Alam h., Hartono R., Ariyoshi K. (2001), *Automatic Summarization of Web Content to Smaller Display Devices*, 6th Int. Conf. on Document Analysis and Recognition, ICDAR01, pages 1064-1068.

NIST web site on summarization. http://wwwnlpirnist.gov/projects/duc/pubs.html, Columbia University Summarization Resources (http://www.cs.columbia.edu/~hjing/ summarization.html) and Okumura-Lab Resources (http://capella.kuee.kyoto-u.ac.jp/index_e.html).

Bighini C., Carbonaro A. (2004), *InLinx: Intelligent Agents for Personalized Classification, Sharing and Recommendation*, International Journal of Computational Intelligence. International Computational Intelligence Society. 2 (1).

Pickens J., Golovchinsky G., Shah C., Qvarfordt P., Back,M, *Algorithmic Mediation for Collaborative Exploratory Search*. To appear in Proceedings of SIGIR.

Freyne J., Smyth B. (2004), *Collaborative Search: Deployment Experiences*, in The 24th SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence. Cambridge, UK, pp. 121-134.

Calic J., Campbell N., Dasiopoulou S., Kompatsiaris Y. (2005), *A Survey on Multimodal Video Representation for Semantic Retrieval*, in the Third International Conference on Computer as a tool, IEEE.

Carbonaro A. (2006), *Defining Personalized Learning Views of Relevant Learning Objects in a Collaborative Bookmark Management System*, In Z. Ma (Ed.), Web-based Intelligent ELearning Systems: Technologies and Applications (pp. 139-155). Hershey, PA: Information Science Publishing.

Burns, Philip R. (2006), *MorphAdorner: Morphological Adorner for English Text*. http://morphadorner.northwestern.edu/morphadorner/textsegmenter/.

Fellbaum C. (1998), *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, Mass.

Bizer C., Lehmann J., Kobilarov G., Auer S., Becker C., Cyganiak R., S. Hellmann (2009), Dbpedia – a crystallization point for the web of data. Web Semantics: Science, Services and Agents on the World Wide Web.

Rafael de Santiago, Andre L.A. Raabe (2010), *Architecture for Learning Objects Sharing among Learning Institutions-LOP2P*, IEEE Transactions on Learning Technologies, pp. 91-95, April-June.

Jones K. S. (2007), *Automatic summarizing: The state of the art. Information Processing and Management*, 43, 1449–1481.

Milne D., Witten I. H. (2008), *An effective, low-cost measure of semantic relatedness obtained from wikipedia links*

Suchanek F. M., Kasneci G., Weikum G. (2008), *Yago: A large ontology from wikipedia and wordnet*.

Yu J., Thom J. A., Tam A. (2007), *Ontology evaluation using wikipedia categories for browsing*.

Kittur A., Chi E. H., Suh. B. (2009), *What's in wikipedia? Mapping topics and conflict using socially annotated category structure*.

Zesch T., Muller C., Gurevych I. (2008), *Extracting lexical semantic knowledge from wikipedia and wiktionary*.

Wolf E., Gurevych I. (2010), *Aligning sense inventories in Wikipedia and wordnet*.

Milne D., Witten I. H. (2009), *An open-source toolkit for mining wikipedia*.

Schonhofen P. (2009), *Identifying document topics using the Wikipedia category network*.

Medelyan O., Witten I. H., Milne D. (2008), *Topic indexing with wikipedia*.

Amiri H., Rahgozar M., Ahmad A. A., Oroumchian F. (2008), *Query expansion using wikipedia concept graph*.

Mihalcea R., Csomai. A. (2007), *Wikify! linking documents to encyclopedic knowledge*.