Aemoo: Exploratory Search based on Knowledge Patterns over the Semantic Web

Alberto Musetti[3], Andrea Nuzzolese[13], Francesco Draicchio[13], Valentina Presutti[3], Eva Blomqvist[23], Aldo Gangemi[3], and Paolo Ciancarini[13]

[1] Dipartimento di Scienze dell'Informazione, Università di Bologna, Italy
[2] IDA, Linköping University, Sweden
[3] STLab-ISTC, Consiglio Nazionale Delle Ricerche, Italy

Abstract. Aemoo is a Web application supporting exploratory search over the Semantic Web. Through a simple keyword-based search interface, users can query Aemoo for information about any entity, which is then collected by aggregating knowledge from diverse sources such as linked data, Wikipedia, Twitter, and Google News. Such aggregation is performed according to cognitively-sound principles through the exploitation of knowledge patterns, and by exploiting semantic relations as well as interpreting hypertext links. Aemoo provides users with an effective summary of knowledge about an entity, including explanations that clarify its relevance, and presents it through a user-friendly interface that supports exploration of further knowledge.

1 Introduction

The Web is a huge source of knowledge and one of the main research challenges is to make such knowledge easily and effectively accessible to Web users. Applications of the Web of data, social networks, news services, search engines, etc., attempt to address this requirement but it is still far from being solved, due to the many challenges arising, e.g. the heterogeneity of sources, using different representations, the implicit semantics of links, as well as the sheer scale of data on the Web.

Although there are several semantic mashup and browsing applications, such as [4, 1, 2], most of them focus mainly on presenting linked data coming from different sources, and visualizing it in interfaces that mirror the linked data structure. Typically, they rely on the semantic relations that are explicitly asserted in the linked datasets, or in explicit annotations, e.g., microdata, without exploiting additional knowledge, e.g. coming from hypertext links, which makes both the data provided and its visualization and navigation quite limited. The problem of delivering tailored and contextualized knowledge remains unsolved as all knowledge that can be retrieved is returned without tailoring it with any context-based rationale.

Other applications focus on text enrichment by performing identity resolution of named entities, and showing associated pictures and news items, such as

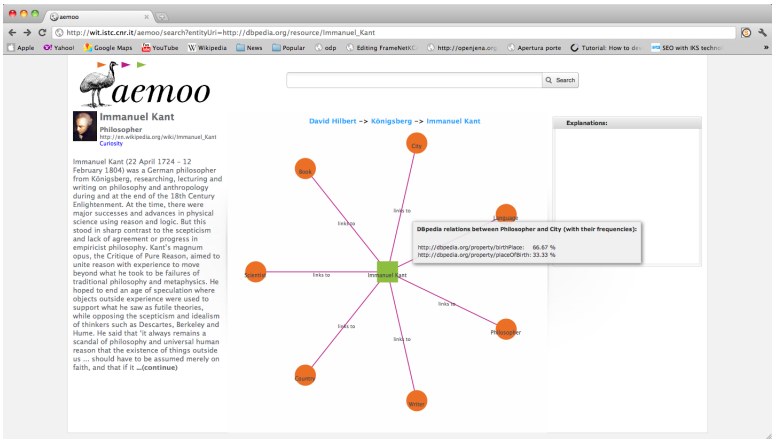

Fig. 1. Aemoo: initial summary page for query "Immanuel Kant".

Zemanta[4] and Calais[5]. However, such applications do not focus on organizing and presenting the discovered knowledge using conceptual-semantic criteria, nor do they try to explain why, e.g., a piece of news, or a set of resources, are relevant with respect to a certain query result, which can be quite confusing for a Web user. Furthermore, to the best of our knowledge no approaches attempt to filter knowledge before presenting it, i.e., drawing a meaningful boundary around the retrieved data in order to limit the result to what is actually meaningful rather than what happens to be linked in the dataset.

In this paper we present **Aemoo**[6], an application that supports exploratory search based on encyclopedic knowledge patterns (EKPs)[7] [3], which exploits Semantic Web techniques and the structure of Web links for enriching query results with relevant related knowledge coming from diverse Web sources. In addition, Aemoo filters the retrieved knowledge, to show a reasonable and relevant set to the user, including the motivation of why a certain piece of information is included.

2 Usage Scenarios

Scenario 1: Knowledge Aggregation and Explanations. Pedro is a high school student, his homework today is to write a report about Immanuel Kant (IK). He types "Immanuel Kant" in the search interface, and Aemoo returns a summary page about him (cf. Figure 1). On the left side of the page, Pedro can read that IK is a philosopher, together with some general information about him, and view a thumbnail image. This information will be enriched as a con-

[4] http://www.zemanta.com/
[5] http://www.opencalais.com/
[6] Aemoo, http://wit.istc.cnr.it/aemoo
[7] An EKP is a special kind of Knowledge Pattern (KP), targeting the description of knowledge entities. Currently about 200 EKPs are available at http://www.ontologydesignpatterns.org/ekp/

sequence of, and during, his navigation. At the same time, a concept map built around IK (as central node) has appeared in the center of the page. The circular nodes signify sets of resources of a certain type (the type is shown as the label on the node), we refer to them as *set nodes*. Those types are the ones that a user would expect to see connected to a Philosopher, i.e., they belong to the Philosopher EKP, according to the empirical study described in [3]. An infotip appearing, when hovering over a link between IK and a set node, shows a list of possible semantic relations that can be assigned to that specific link, according to their frequencies in DBpedia. For example, the relations between IK and cities could be `birthPlace` or `placeOfBirth` if we considered only linked data asserted relations. This example shows the limitation of such an approach, in fact cities can be related to Immanuel Kant also for other reasons than being his place of birth, which can be explained in Aemoo by additional information in the explanations window (right side of the interface). An infotip that appears when hovering with the cursor over a set node shows the list of resources contained in the set, meaning that those resources are for some reason connected to IK. For example, IK links to a set of scientists, which is interesting information for Pedro. Hence, he clicks on the set node *Scientist*, because he wants to know more about those particular relations. The graph moves and re-centers around the relation between IK and the set of scientists (cf. Figure 2); Aemoo narrows its context to IK's relations with scientists. The set node *Scientist* has also "exploded", and all its members are visible. When Pedro holds the mouse cursor over David Hilbert, the summary page of IK is on the right side enriched by text paragraphs that explain why IK is linked to David Hilbert. This information comes from DBpedia, as well as other sources such as Twitter and the Google news service. For example, Pedro notices that "David Hume" is included in the set of philosophers related to IK. This resource was added to Aemoo's IK map because the application found a news article that relates the two philosophers. If Pedro hovers his cursor over David Hume, an extract of the news article is shown as the explanation of this particular relation.

Scenario 2: Exploratory search. Pedro, however, would like to collect some more information about David Hume, hence, he clicks on that resource, and Aemoo changes its context from IK to David Hume, showing a summary page for him instead. Pedro can continue to navigate, by clicking nodes of the graph map, to explore also this entity and its related information. Through the breadcrumb (located between the search field and the graph map (cf. Figure 2) Pedro can go back and forth, and revisit his navigation path and the associated knowledge.

Scenario 3: Curiosity. Eva is a member of the editorial board of a TV program that dedicates each episode to a different country. Now she has to edit the episode about Italy. She uses Aemoo as described above, for building a summary about the country that can be useful for the introductory part of the show. However, then she wants to find additional information that is special, or "peculiar", i.e. non-obvious, for making the episode more interesting to her audience. Hence, she clicks on the "curiosity" link (cf. Figure 2), which causes Aemoo to change the perspective of the summary it builds about Italy. In fact, she is presented

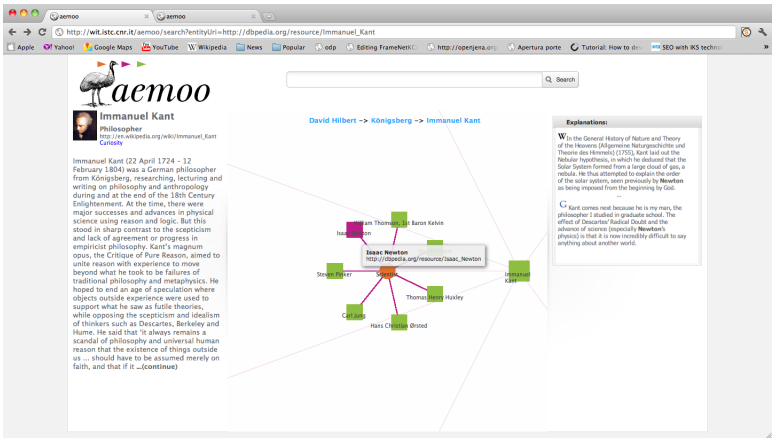

Fig. 2. Aemoo: browsing relations between "Immanuel Kant" and scientists. Explanations, breadcrumb and curiosity.

with additional knowledge about Italy, which was not previously included in the summary due to the filtering through the EKP for countries. What is now shown are "special" facts about Italy, things that are not usually used to describe a country. The knowledge is again visualized as a graph map, and enriched with news and tweets just as before, but this time the set nodes are selected with a different criterion: they are types of resources that are unusual to be included in the description of a country, hence possibly peculiar.

3 Semantics in Aemoo

This section describes the *Aemoo approach*, with particular focus on how the application exploits semantics to perform its task and to improve the user experience.

Aemoo exploits DBpedia as a starting point, e.g. for identity resolution, to resolve a user query, but then proceeds to enrich the DBpedia data with additional information coming from other sources, such as Wikipedia (in particular its link structure), Google news, and Twitter. Aemoo performs *identity resolution* in two main situations: (i) it identifies the identity of the resource referred to by a user query, and (ii) it identifies the identity of the resources that are mentioned in news and tweets together with the subject of the exploration. For example, if a user is exploring knowledge about Steve Jobs, and there is a tweet about "Steve Jobs leaving his place at Apple to Tim Cook", Aemoo will recognize Tim Cook and Apple and would add them to the set of entities related to Steve Jobs.

Secondly, Aemoo recognizes resource types according to the DBpedia taxonomy of types. The type is important because (i) it provides users with additional information about an entity, and (ii) it affects the presentation of knowledge, which is based on knowledge patterns (KPs). A specific kind of KPs, Encyclopedic KPs (EKPs), have been shown to provide a *cognitively-sound organization*

and selection of information for a resource type; see the experimental user-study on encyclopedic knowledge patterns (EKP) from Wikipedia links in [3]. In other words, each DBpedia type is associated with an EKP, and after identifying the resource type the associated EKP can be retrieved. The EKP is then used as a query for enriching the knowledge concerning the resource of interest, but also for deciding on what information to present to the user. The resulting knowledge structure is presented using a user-friendly graphical interface, i.e. a *graph map*, where user actions are associated with events that further specify or enrich the result context. EKPs are used also for identifying what we call "curiosity". In fact, Aemoo exploits long-tail information from EKPs for building a different perspective over the knowledge related to an entity, which include peculiar facts instead of more expected and encyclopedic descriptive information.

Another important aspect is that Aemoo attempts to interpret the semantics of Web links, in particular Wikipedia links. The EKPs [3] mentioned above have been extracted from the graph of Wikipedia links available in the DBpedia dataset. By using those EKPs Aemoo provides the users with useful information that can explain the implicit semantics of such untyped relations: (i) each link between a resource and its associated set of nodes is annotated with a list of semantic relations, and their frequencies, that occur in DBpedia between that resource type and the type of the set node. This information is visible through an infotip associated with the links in the graphical visualization of the EKP instances. Additionally, (ii) the text surrounding the link is visualized in the *explanation* box, which gives a contextual description of the relation.

Aemoo results are highly contextualized. The context is given by an EKP, determined based on the resource type. All relations between resources that emerge from the selected EKP, and those emerging from unstructured text, i.e. tweets and news, are used as the basis for selecting the information to be aggregated and visualized. The main knowledge context of a user is always clearly visible in the interface, it is explicitly described on the left hand side through the name of the resource under investigation and descriptive text. As the context changes, e.g. through user navigation, the description changes accordingly, helping to avoid disorientation and confusion. Furthermore, the visualized information is also tailored based on user interactions, e.g. mouse roll-over on a related entity shows an explanation of its relation to the main context (main resource being investigated).

In summary, Aemoo can be said to perform KP-based exploratory search, and this is what makes Aemoo especially novel. It exploits the structure of linked data, and organizes it by means of EKPs for supporting exploratory search. The use of EKPs allows to draw meaningful boundaries around data, i.e. tell us what links to follow and which not to bother with. In this way, Aemoo performs both *enrichment* and *filtering* of information, based on the structure of the EKP, which in turn reflects the most common way to describe entities of this particular type [3]. Hence, the user benefits from such presentation and is guided in the navigation. Instead of navigating a plain graph of triples she navigates *units of knowledge* and moves from one to the other without loosing the overview of an entity.

4 Implementation

Aemoo is released as a Web application composed by a server side component based on a REST Web service in Java, and a client side component based on HTML and JavaScript. The client side interacts with third party components via REST interfaces through AJAX.

The server side exposes a REST service for retrieving EKP-based graphs as well as "curiosity graphs" about entities. Its input is an entity URI, e.g. dbpedia:Barack_Obama. Its output is an RDF graph corresponding to the summarization provided by the entity's EKP (selected based on the entity type). The RDF graph is obtained by generating a SPARQL CONSTRUCT out of the selected EKP. EKPs are available as OWL ontologies[8], and there are currently around 200 EKPs that have been extracted from DBpedia and the Wikipedia linking structure (cf. [3]).

The client side component handles the graphical visualization of Aemoo through the JavaScript InfoVis Toolkit[9], a Javascript library that supports the creation of interactive data visualizations for the Web. Abstracts and thumbnails are retrieved by querying the DBpedia SPARQL endpoint exposed as a REST service.

Aemoo also detects relations between the investigated entity and other entities from Twitter [10] as well as from Google News[11]. Tweets and news are retrieved using their respective REST services. For performing identity resolution on user queries, tweets, and news we use Apache Stanbol[12]. Entities recognized in tweets and news are dynamically added to the graph map. The textual context around a wikilink from Wikipedia, tweets, and extracts of news are used for explaining relations between entities. The explanations are associated with provenance information (by visualizing a logo or the name of the source before the text). Currently, the used sources are "'Wikipedia", "Twitter" and any news publisher, e.g. "Reuters", available in Google News.

5 Future Challenges

Future plans include to further improve Aemoo, and use it as a test-bed for our research work. The main research challenges we are currently working on include: (i) Empowering automatic interpretation of hypertext links and text in general, by using NLP techniques, e.g., frame detection, combined with knowledge patterns and linked data analysis, for making Aemoo able to recognize situations, and allow users to automatically annotate interesting text. (ii) Improving explanations by more extensively exploiting linked data and in particular its formal semantics, available behind the visualized knowledge. This data can come from datasets as well as from recommending strategies, frame detection, knowledge patterns, etc. (iii) Allowing users to investigate different snapshots in time, rep-

[8] http://www.ontologydesignpatterns.org/ekp/
[9] http://thejit.org/
[10] https://search.twitter.com/search.json
[11] https://ajax.googleapis.com/ajax/services/search/news
[12] http://incubator.apache.org/stanbol/

resenting the same resource, i.e., comparing the information about an entity in several different periods, e.g., Barack Obama before he was president of the United States and today. (iv) Allowing users to extract a customized map out of the Aemoo summary (as an extension of the breadcrumb feature), and to edit the labels of the relations. This would allow us to also monitor what semantics is provided by users, with respect to the links between entities. (v) Finally, we want to improve the graphical user interface of Aemoo, by first conducting a rigorous user-based evaluation in order to find the most appropriate directions of improvement.

Acknowledgements This work has been part-funded by the European Commission under grant agreement FP7-ICT-2007-3/ No. 231527 (IKS - Interactive Knowledge Stack)

A Addressing Semantic Web Challenge Requirements

Table 1 describes how Aemoo addresses minimal requirement specified by the Semantic Web Challenge call. Aemoo addresses also a number of desirable features which are described in Table 2.

The application has to be an end-user application, i.e. an application that provides a practical value to general Web users or, if this is not the case, at least to domain experts.
Aemoo is an *end-user application*, targeting the general use case of finding and exploring information about entities on the Semantic Web. The main contributions, i.e. novelty and value for end users, are its ability to filter and extend (integrate) information about a certain entity based on cognitively sound knowledge patterns, as well as to allow users to navigate and explore that knowledge in an intuitive manner.
The information sources used should be under diverse ownership or control, should be heterogeneous (syntactically, structurally, and semantically), and should contain substantial quantities of real world data (i.e. not toy examples).
In order to provide information to the user, Aemoo integrates information from *large-scale heterogeneous sources*, both semantically represented, e.g. DBpedia, and textual sources, e.g. Tweets. Worth pointing out is also that several of the sources are highly *dynamic*, i.e. Tweets and news items arrive in a continuous "stream" of data.
The meaning of data has to play a central role. Meaning must be represented using Semantic Web technologies. Data must be manipulated/processed in interesting ways to derive useful information and this semantic information processing has to play a central role in achieving things that alternative technologies cannot do as well, or at all;
Aemoo uses both NLP techniques such as identity resolution, and Semantic Web technologies such as knowledge patterns, and linked data in order to represent the *meaning of data*. Aemoo also identifies entity types and use them for selecting the most appropriate knowledge pattern, which is in turn used for expressing and selecting the knowledge to be visualized. While identity resolution is not extremely novel, *the use of knowledge patterns to filter, extend and visualize knowledge is completely new*. Knowledge patterns can be viewed as expressing the nature of a certain type of entity, i.e. expressing its semantic context.

Table 1. Addressing Minimal Requirements

References

1. Heim, P., Hellmann, S., Lehmann, J., Lohmann, S., Stegemann, T.: RelFinder: Revealing relationships in RDF knowledge bases. In: Proceedings of the 3rd International Conference on Semantic and Media Technologies (SAMT). Lecture Notes in Computer Science, vol. 5887, pp. 182–187. Springer (2009)

2. Heim, P., Ziegler, J., Lohmann, S.: gFacet: A browser for the web of data. In: Auer, S., Dietzold, S., Lohmann, S., Ziegler, J. (eds.) Proceedings of the International Workshop on Interacting with Multimedia Content in the Social Semantic Web (IMC-SSW'08). pp. 49–58. CEUR-WS (2008), http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-417/paper5.pdf
3. Nuzzolese, A.G., Gangemi, A., Presutti, V., Ciancarini, P.: Encyclopedic knowledge patterns from wikipedia links. In: Aroyo, L., Noy, N., Welty, C. (eds.) Proceedings fo the 10th International Semantic Web Conference (ISWC2011). pp. 520–536. Springer (2011)
4. Tummarello, G., Cyganiak, R., Catasta, M., Danielczyk, S., Delbru, R., Decker, S.: Sig.ma: Live views on the web of data. Web Semantics: Science, Services and Agents on the World Wide Web 8(4), 355 – 364 (2010), http://www.sciencedirect.com/science/article/pii/S1570826810000624, semantic Web Challenge 2009, User Interaction in Semantic Web research

The application provides an attractive and functional Web interface (for human users)
Aemoo has an *attractive user interface*. The mode of interaction lets users navigate the knowledge based on the structure of EKPs. This is a novel way of interacting with semantic information. In addition, the interface seamlessly combines structured and unstructured information, i.e. users are presented with both textual information and graphical visualizations of data at the same time.
The application should be scalable (in terms of the amount of data used and in terms of distributed components working together). Ideally, the application should use all data that is currently published on the Semantic Web. Rigorous evaluations have taken place that demonstrate the benefits of semantic technologies, or validate the results obtained.
Aemoo uses a number of diverse information sources, all large and domain-independent. Some of them are in addition dynamically growing. Although no experiments have been conducted concerning the application scalability, the fact that it *already uses diverse large-scale resources* indicates a robustness and scalability of the approach. The application is also extensible, i.e. although the application at the moment does not use the "complete Semantic Web" it is open to be easily extended with more sources. A rigorous evaluation has been conducted on the effectiveness of EKPs, which proved to be cognitively-sound components for organizing knowledge [3]
Novelty, in applying semantic technology to a domain or task that have not been considered before
The Aemoo approach is *highly novel*, since to the best of our knowledge it is the only Semantic Web application so far that applies explicit representations of knowledge patterns i.e. EKPs, based in human cognition, to organize, present, and guide exploration of knowledge to a user.
Functionality is different from or goes beyond pure information retrieval
While information retrieval classically performs retrieval of documents, and leaves the actual extraction of information to the user, our approach is in line with the vision of the Semantic Web by instead directly retrieving pieces of knowledge. The use of EKPs also *goes beyond what is usually expected from an information retrieval system*, since they facilitate both extension and filtering of the retrieved information, as well as integration of information from several sources.
The results should be as accurate as possible (e.g. use a ranking of results according to context)
Since the Aemoo's task is to retrieve relevant information describing an entity, measuring *accuracy* means to measure how well the retrieved information describes the entity in question. Such an experiment was reported in [3], and shows that EKPs are indeed intuitive and relevant as ways of retrieving and organizing knowledge.

Table 2. Addressing Additional Desirable Feature