Ontos Clip and Share

Philip Dudchuk
Text Mining Group
AviComp ZAO
Moscow, Russian Federation
philip.dudchuk@avicomp.ru

Victor P. Klintsov
Department of Technology
Ontos AG
Biel, Switzerland
victor.klintsov@ontos.com

Daniel Hladky
Department of Business
Ontos AG
Biel, Switzerland
daniel.hladky@ontos.com

ABSTRACT

This project describes an application for creating ubiquitous hypertext on the Web, which enhances the user experience by allowing clipping and sharing the information. The goal of the application is to annotate text and link it to relevant content, especially from the Linked Open Data (LOD) community and from the Ontos knowledge base. The paper describes two use cases and highlights the main functionality of the application.

Categories and Subject Descriptors

H.3.1 [**Information Systems**]: Information Storage and Retrieval

General Terms

Semantics.

Keywords

RDFa, Annotation, Hypertext, Knowledge Sharing, Clipping.

1. INTRODUCTION

Ontos Clip and Share (OCaS) is an application that automatically injects hypertext to pieces of text content the user faces on the Web. It employs DBpedia as a controlled vocabulary, Ontos Semantic Knowledge Base [Hladky 2010] as the data integration instrument, and RDFa as the standard format for embedding semantic metadata into HTML documents.

OCaS enables two major functions:

1. Providing relevant interlinked content for any web page the user visits.
2. Clipping text pieces from different pages into a single document and sharing it through e-mail, in the user's blog, social services, etc. The resulting document (clip) also contains hypertext and RDFa tags.

OCaS is implemented as an add-on for the top web browsers. Once the user installs it, DBpedia objects become highlighted on every web page the user visits. When the user hovers mouse over a highlighted item, a special balloon pops up which displays content from Ontos Semantic Knowledge Base which is relevant to the entity and to the context in which it occurs. For instance, in a news article about the crisis in Greece, a highlighted occurrence of Germany will be linked to other articles exactly about the Germany linked to the crisis in Greece, but not to everything about Germany.

Furthermore, through the same balloon (see Figure 2) the user gets a link to the corresponding Wikipedia entry, as well as multimedia content from YouTube, Flickr and other major web services depending on the type of the highlighted item.

2. Technology behind the application

OCaS is based on the proprietary technology of semantic indexing of web pages with adoption of DBpedia entries as controlled vocabulary for entity extraction.

First, the semantic engine extracts named entities and facts/events which interlink the entities in the document [Efimenko 2010]. The engine also defines most significant entities for the web-page. The significance is calculated as a complex measure of the position of the occurrence in the text, the overall number of occurrences and the number of connected events and facts extracted from the document.

Then, to display most relevant data in the balloon, the system forms a search query about the entity in question and the connected facts, events and most significant entities in the document. The relevant web services including Flickr, YouTube and Ontos return links to pictures, video, documents, etc.

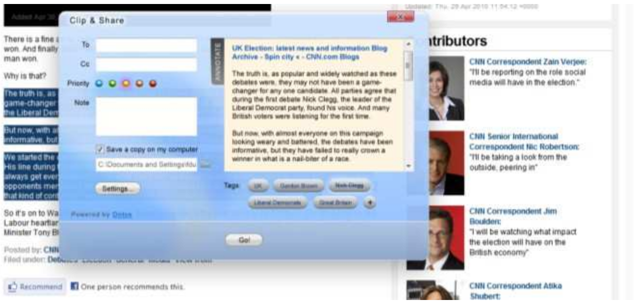
Ontos semantic knowledge base (Ontos KB) currently contains RDF-grounded semantic indexes for entities, facts and events extracted from 5+ million documents mainly from Russian and English online news. The whole machinery resides on a scalable grid infrastructure, which allows extending the overall collection of indexed documents permanently.

3. Use Cases

3.1 Benefits for end users

With the help of OCaS add-on, the user can clip pieces of different web-pages into a single document (see Figure 1) which also contains semantic annotations based on RDFa vocabs, and share it with people via e-mail, blogs and other social services. Furthermore, the system suggests recommended tags for the selected text.

I-SEMANTICS 2010, September 1–3, 2010, Graz, Austria.
Copyright © 2010 ACM 978-1-4503-0014-8/10/09... $10.00.



There is a fine
won. And finally
man won.

Why is that?

The truth is, as
game-changer
the Liberal Dem

But now, with al
informative, but

We started the
His line during
always get ever
opponents mer
that kind of con

So it's on to Wa
Labour heartlar
Minister Tony Bl

Posted by: CNN
Filed under: Deb

One person recommends this.

UK Election: latest news and information Blog Archive - Spin city « - CNN.com Blogs

The truth is, as popular and widely watched as these debates were, they may not have been a game-changer for any one candidate. All parties agree that during the first debate Nick Clegg, the leader of the Liberal Democrat party, found his voice. And many British voters were listening for the first time.

But now, with almost everyone on this campaign looking weary and battered, the debates have been informative, but they have failed to really crown a winner in what is a nail-biter of a race.

CNN Correspondent Zain Verjee:
"I'll be reporting on the role social media will have in the election."

CNN Senior International Correspondent Nic Robertson:
"I'll be taking a look from the outside, peering in"

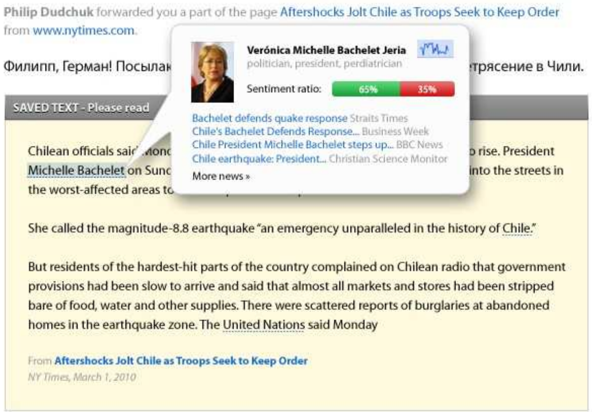
CNN Correspondent Jim Boulden:
"I will be watching what impact the election will have on the British economy"

CNN Correspondent Atika Shubert:

Figure 1: Clipping text fragment

With OCaS the user gets the functions that allow to:

- get semantic hypertext on any web-page linking a named entity to most relevant data
- select pieces of text, clip it into a single document, which contains semantic hypertext
- share the clipped data via e-mail, social services such as BlogSpot, LiveJournal, LiveInternet, social networks, social bookmarks, etc.



Philip Dudchuk forwarded you a part of the page Aftershocks Jolt Chile as Troops Seek to Keep Order from www.nytimes.com.

Филипп, Герман! Посылак трясение в Чили.

Verónica Michelle Bachelet Jeria
politician, president, perdiatrician

Bachelet defends quake response Straits Times
Chile's Bachelet Defends Response... Business Week
Chile President Michelle Bachelet steps up... BBC News
Chile earthquake: President... Christian Science Monitor

Chilean officials saic Mond o rise. President Michelle Bachelet on Sunc into the streets in the worst-affected areas to

She called the magnitude-8.8 earthquake "an emergency unparalleled in the history of Chile."

But residents of the hardest-hit parts of the country complained on Chilean radio that government provisions had been slow to arrive and said that almost all markets and stores had been stripped bare of food, water and other supplies. There were scattered reports of burglaries at abandoned homes in the earthquake zone. The United Nations said Monday

From Aftershocks Jolt Chile as Troops Seek to Keep Order
NY Times, March 1, 2010

Figure 2: Automatic marked Named Entity with enriched data from LOD and Ontos KB

At the end of the day, the user gets a possibility to avoid the routine of copying pieces of text corresponding to named entities and searching relevant information in popular services. When hovering over a recognized named entity the user gets most relevant information and has then the possibility to navigate into more information by following the link (see Figure 2).

Bloggers get an additional benefit due to the fact that the content shared via OCaS gets RDFa tags with both DBpedia and Ontos KB IDs. All major search engines (Google, Bing, Yahoo) have recently adopted the RDF technology, and now documents with embedded RDFa get a higher page rank [Herman2009].

3.2 Benefits for content holders

The holders of large volumes of content, primarily online mass media also benefit from adoption of the proposed technology. Media corporations get a possibility to offer semantic hypertext within their own add-ons. As soon as they do so, they will have an effective instrument for targeting users by their habit and profile and link more naturally relevant content. This will allow keeping users longer on the web page and providing a possibility to offer new services [Lovinger 2010].

Imagine for instance a toolbar of a certain media agency that displays news online feed. The user, who already installed the toolbar, is currently browsing news on a competing media portal, and is exploring a story, say, about the eruption of the volcano in Iceland. At this very moment the toolbar should display the content related to the volcano, and attract the user to the toolbar holder's story and multimedia content. This is quite possible through the on-the-spot analysis of the entities and their links on the viewed web page and filtering the toolbar's content by the corresponding query. Such targeting strategy can be also extended to ad content.

Furthermore, the content clipped by the user can also be automatically analyzed. Entities found in the user's clip and some other closely related entities form the scope of the user's current interest. Consequently, the content on the news portal can be customized automatically to take into account the user's interests.

On blog hosting portals and in social networks the analysis of user's interests with the help of the proposed technology can be performed on user friend clusters gathered around certain topics.

4. Future Work

Since the LOD community is growing we will focus in the near future on augmenting the possibility to interlink other sources. Further work is devoted to the area of co-referencing [Glaser 2009], for example by using OKKAM[1]. Another important aspect is the personalization of the annotation process which will allow the user to define own vocabularies, especially for named entity recognition.

5. REFERENCES

[1] Efimenko I. et al 2010. Providing Semantic Content for the Next Generation Web. Book Chapter of "Semantic Web", INTECH, ISBN 978-953-7619-54-1.

[2] Herman I. 2009. RDFa usage spreading. http://ivan-herman.name/2009/12/12/rdfa-usage-spreading%E2%80%A6/.

[3] Hladky, D. 2010. Sustainable Advantage for the Investor Relations Team through Semantic Content. Book Chapter of "Semantic Web", INTECH, ISBN 978-953-7619-54-1.

[4] Glaser, H., Jaffri, A. and Millard, I. 2009. Managing Co-reference on the Semantic Web. In: WWW2009 Workshop: Linked Data on the Web (LDOW2009), 20 April 2009, Madrid, Spain.

[5] Lovinger, R. 2010. Nimble: a razorfish report on publishing in the digital age. http://nimble.razorfish.com/

[1] http://www.okkam.org/