NATURAL LANGUAGE PROCESSING AND WEB MINING

APPLICATION OF SOCIAL ANALYTICS FOR BUSINESS INFORMATION SYSTEMS

Alexander Troussov, D.J. McCloskey

***Abstract**: Social networking tools, blogs and microblogs, user-generated content sites, discussion groups, problem reporting, and other social services have transformed the way people communicate and consume information. Yet managing this information is still a very onerous activity for both the consumer and the provider, the information itself remains passive. Traditional methods of keyword extraction from text based on predefined codified knowledge are not well suited for use in such empirical environments, and as such do little to support making this information more an active part of the processes to which it may otherwise belong. In this paper we analyse various use cases of real-time context-sensitive keyword detection methods using IBM LanguageWare applications as example. We present a general high-performance method for exploiting ontologies to automatically generate semantic metadata for text assets, and demonstrate examples of how this method can be implemented to bring commercial and social benefits. In particular, we overview metadata-driven semantic publishing on the BBC FIFA World Cup 2010 website and the applications for social semantic desktops.*

***Keywords**: data mining, natural language processing, recommender systems, social semantic web, graph-based methods.*

***ACM Classification Keywords**: H.3.4 [Information Storage and Retrieval]: Systems and Software – information networks; H.3.5 [Information Storage and Retrieval]: Online Information Services – data sharing.*

Introduction

The massive scale production of human oriented information presents individuals and organizations with serious problems imposed by the limitations of the HTML and associated text based knowledge containers ubiquitous on the web. In this paper we

examine the general problem through study of a concrete example for each case, challenges facing the individual information consumer, the activity centric environment Nepomuk-Simple, and the information centric organization, BBC's World Cup 2010 website.

The core focus of a leading news and media organization is on providing the best and most up to date information in the most consumable way for the expected audience. The amount of data now flowing through our electronic world means that even in a controlled environment where a fixed number of journalists report on a topic there is still a need to relate and integrate the information created by those journalists with the torrential flows from all other channels. The definition of the "best and most up to date information" has shifted radically since the advent of the internet and accelerated through the web 2.0 social media explosion. Similarly, the definition of "consumable" and the nature of the audience have made step changes with advances in recommender systems, web presentation and information visualization and the expectations of automated feed consumption and programmatic access for mashup and situational or application based re-integration of the information content. In this paper we show that for many of the most pressing issues the semantic web standards coupled with state of the art text analysis techniques can make it possible for the information we place on the web to become programmatically available.

Converting the textual output of the social web into programmatically accessible knowledge is a well recognized challenge today. The components of a successful strategy in addressing this challenge are, at a high level:

1. Text analysis – for extracting the spans of text which represent agents and situations or concepts and relationships.
2. A domain knowledge description system and language – to formally describe the semantics of the real world topics discussed in the human language text.
3. A knowledge base used to store both the domain description and the instance of facts discovered in the text and support further global analysis and query.
4. Semantic analysis – to apply logic and graph based methods to the knowledge graph stored in the knowledge base.

The BBC FIFA World Cup 2010 website in an exemplar in terms of standard web based presentation of multifaceted information. It presents general reference information about the event itself, such as details of teams, fixtures, venues, individual player profiles and also dynamic information such as up to date news and special features, commentator blogs, match reports, video links, and interesting statistics as the tournament progresses. The site is laid out in a simple and classic way: layered menus at the top, blogs down

the left, main pane featuring the focus document or story with related information links to the right and bottom. The challenge was to automate the aggregation of related content on this multidimensional canvass. Manual selection, curation and placement of such information had been the norm, this was hugely labour intensive and would not scale to the new levels of dynamic information creation at such an event.

The rest of the paper is organized as follows. In section 2 we provide an overview of the requirements for natural language processing (NLP) covering the use cases considered in this paper. We show how these requirements were taken into account in IBM LanguageWare tools.

In section 3 we outline the architecture of the hybrid recommender system in the activity centric environment Nepomuk-Simple. In section 4 we overview metadata-driven semantic publishing on the BBC FIFA World Cup 2010 website. Finally, section 5 describes the conclusions and future work.

Dynamic Semantic Tagging

In this section we describe the requirements for NLP covering the use cases considered in this paper, these requirements effectively determine our use of the term dynamic semantic tagging. We show how these requirements were taken into account in IBM LanguageWare tools.

2.1 Requirements

First of all, the goals which drive the application of NLP to a real world problem are often determined by the business model, increasingly these business models are represented by ontology. Put simply, the most important task for the NLP application is to determine how the terms mentioned in a text are related to business model, including term disambiguation and ranking the relevancy of the text to the concept from the semantic network of ontological concepts. The social context, modeled as a network, is a natural extension of the semantic networks which are formed from concepts represented in ontologies. It is possible to use such networks for knowledge based text processing. In fact, layering these different networks upon each other brings mutual benefits, since more contextual linkage becomes available between concepts which might otherwise have been isolated. The gains of additional context come at a price of a challenge to runtime performance of the software which will analyze the network. Furthermore, many use cases demand near real time rapid response from the analytics algorithms and scalability is a major concern. There are also non-functional requirements from the software engineering perspective such easy software maintenance and reuse.

2.1 Outline of the Procedure

The basic procedure of mining texts using graph-based methods has two major steps, as depicted on the Fig. 1: lexical analysis combined with mapping from text to ontology, reasoning how concepts mentioned in a text sit together.

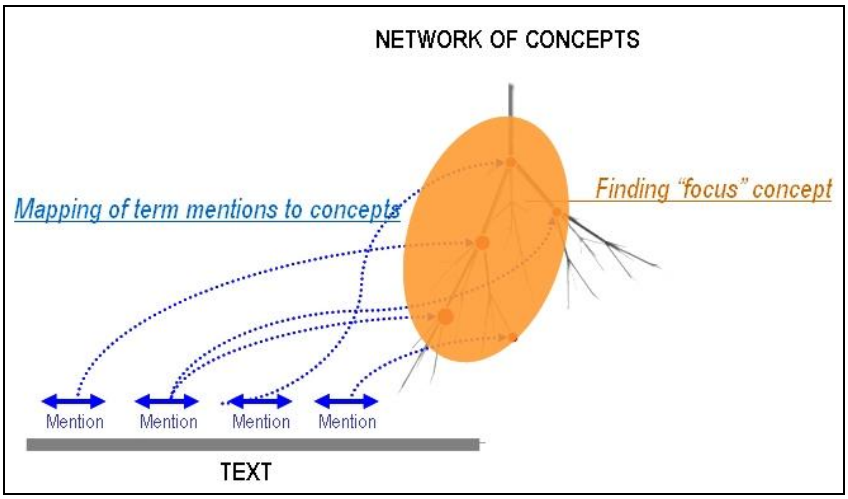

Fig. 1. Procedure of mining semantic models of texts consists of mapping from text to the ontology and reasoning how concepts mentioned in a text sit together with the objective of finding focus concepts.

Correspondingly, in the rest of this section we describe the approach taken by IBM LanguageWare to organise lexico-semantic resources for the purpose of mapping from text to ontology and for the selection of "reasoning" methods suitable for large networks.

2.3 Layered Dictionary Layout and Technical Organisation of Lexico-Semantic Resources into Lexically Enriched Ontology

According to [Ou et al., 2006] "It is an established fact that knowledge plays a vital role in the comprehension and production of discourse. The process of interpreting words, sentences, and the whole discourse involves an enormous amount of knowledge which forms our background and contextual awareness. However, how various types of knowledge can be organized in a computer system and applied to the comprehension process is still a major challenge for semantic interpretation systems used in natural language processing".

In LanguageWare a clear decision was made to separate lexico-semantic resource into two layers – lexical and semantic, and to provide a flexible framework for identifying term mentions of ontological concepts in raw text. Our pragmatic approach for representation of lexico-semantic resources is in vein with the most cited (in Computer Science and in

the field of Artificial Intelligence) Gruber's definition of ontology: "An ontology is an explicit specification of a conceptualization." [Gruber, 1993] with Guarino's clarifications [Guarino, 1998]: "… engineering artefact, constituted by a specific vocabulary used to describe a certain reality…". However, our goal is to design the means by which various types of knowledge can be organized in a computer system and applied to text processing for semantic annotation and IR applications.

The Fig. 2 shows IBM LanguageWare layered organisation of lexico-semantic knowledge.

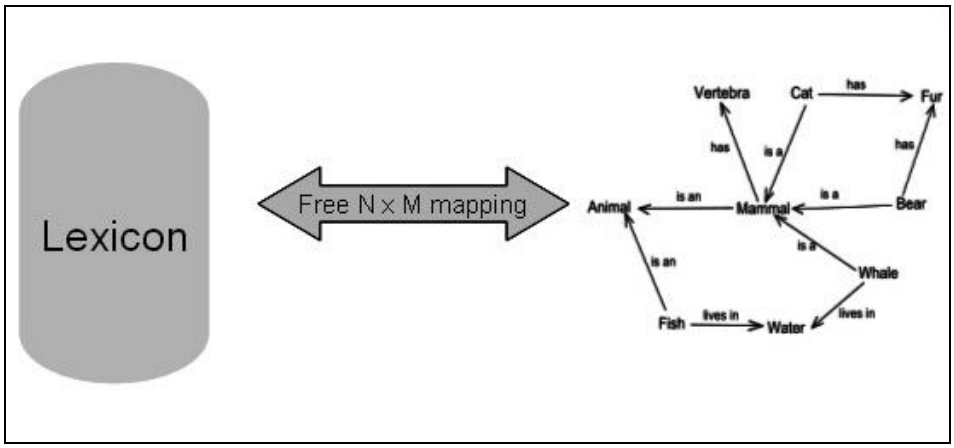

Fig. 2. Layered organisation of lexico-semantic knowledge for automatic text processing: lexical layer, semantic network layer, free N×M mapping between these two layers, lexical ambiguity phenomena (like synonymy and polysemy) are expressed by mapping

The Fig. 3 outlines processing resources which utilise these layers.

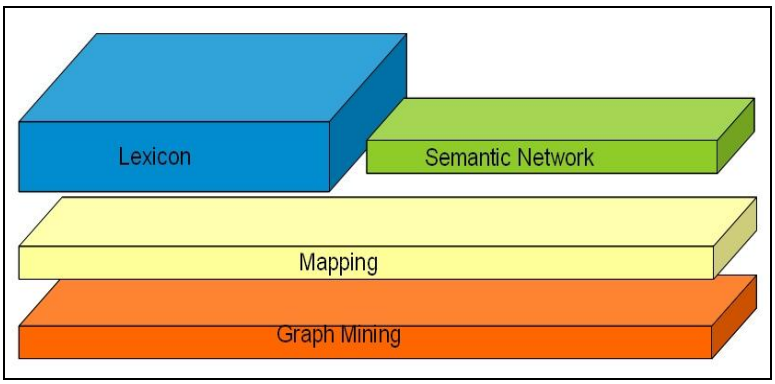

Fig. 3. Layered organisation and processing resources: Lexicon is used by lexical analyser to find mentions of concepts represented by nodes in the semantic network; mapping from text to concepts creates semantic model of a text (as a function on nodes of the network which shows how concepts are -related to text); graph-mining provides analytics on term mentions

2.4 Natural Language Understanding and Graph-based methods

From computational point of view natural language understanding could be considered as inferencing. For instance, based on a taxonomy of geographical locations one might infer that a text which mentions Malahide might be relevant to Canada since Malahide is a township in Elgin County, Ontario, Canada. However, terms are ambiguous (for instance, there is a location Malahide, Co. Dublin, Ireland), the knowledge encoded in ontologies is never "the truth, the whole truth, and nothing but the truth", and high precision mapping from lexical entries to concepts (that is, for instance, to detect that the hotel Paradis Gisenyi Malahide in Rwanda is not related to Canada or Ireland) is first of all prohibitively slow, and secondly, requires creation of hand-coded rules specific to the ontology in question.

At the same time, if we analyze how the concepts mentioned in a text sit together, one can reason that a text which mentions Malahide and Europe – is a little bit more likely to be about Ireland than about Canada, text which mentions Malahide and Clontarf – is more likely to be about Ireland than about Canada, and cohesive coherent text which mentions: Malahide, Mulhuddart, Lansdowne, Clontarf, Donabate (all these locations are in the capital of Rep. of Ireland) - is almost for sure about Dublin.

This type of "fuzzy" inferencing could be efficiently implemented using methods of "soft mathematics", in our case – graph-based methods.

Formally, solution of many network data mining tasks boils down to the following problem: given an initial function $F_0(v)$ on the network nodes, construct the function $F_{lim}(v)$ which provides the answer. In different domains the function F_0 could be referred to as the initial conditions, the initial activation, semantic model of a text, etc. In ontology based text processing, the initial function F_0 is the semantic model of a text w.r.t. to the knowledge: for instance, $F_0(v)=0$ if the concept v is not mentioned in the text, $F_0(v)=n$ if the concept v is mentioned *n* times. The function $F_{lim}(v)$ should show the foci of the text; for instance, $Argmax(F_{lim})$ is the most important focus of the text, while $F_{lim}(Argmax(F_{lim}))$ is the numerical value of the "relevancy". In information retrieval, the link analysis (such as Google's PageRank ([Brin and Page, 1998], [Langville and Meyer, 2006]) ranks web pages based on the global topology of the network by computing $F_{lim}(v)$ using the iterative procedure where the initial condition is that all web pages are equally "important" ($F_0(v) \equiv 1$).

Computationally efficient and scalable algorithms usually compute the function F_{lim} using iterations: on each iteration the value of $F_{n+1}(v)$ is computed depending on the values of the function F_n on the nodes connected to the node v. This is a very broad range of

algorithms including PageRank, spreading activation, computation of eigenvector centrality using the adjacency matrix.

Most of the mathematical algorithms behind such iterative computations are the "network flow" algorithms: they are based on the idea that something is flowing between the nodes across the links, and the structural prominence of nodes could be explained and computed in terms of incoming, outgoing and passing through traffic. Similar iterative computational schemes have been used for long time in finite element analysis to solve physical problems including propagation of heat, of mechanical tensions, oscillations, etc. Although finite element analysis automata usually perform on rectangular (cubic, etc.) grids, the extension to arbitrary networks is feasible [Troussov et al, 2011a].

In the IBM LanguageWare products described in this paper, we exploited the principal approach based on the computational scheme described above. The rationale for our choice of the graph-based method, which could be described as network flow method, could be understood by comparison with the mature area of applications of graph-based methods such as social network analysis. The task of determining "important" concepts based on the analysis of how the concepts mentioned in a text sit together, is a task of finding structurally important nodes in the network of concepts. In social network analysis, many traditional measurements of structural importance could be viewed in accordance with dynamic model-based view of centrality in [Borgatti, 2005] "that focuses on the outcomes for nodes in a network where something is flowing from node to node across the edges" [Borgatti and Everett, 2006].

Specific version of the network flow method used was spreading activation, which allows us to provide further optimisation of the performance by using bread-first search ([Troussov et al., 2009]). In addition to this principal graph-mining technique, we also exploited additional empirics, such as one sense per discourse [Judge et al., 2008].

Hybrid Recommender System in the Activity Centric Environment Nepomuk-Simple

In this section we outline the architecture of the hybrid recommender system in the activity centric environment Nepomuk-Simple (EU 6th Framework Project NEPOMUK) following [Troussov et al., 2008].

"Real" desktops usually have piles of things on them where the users (consciously or unconsciously) group together items which are related to each other or to a task. The "pile" based graphical user interface, used in the Nepomuk-Simple, imitates this type of data and metadata organisation which helps to avoid premature categorisation and reduces the retention of useless documents.

Metadata describing user data is stored in the Nepomuk Personal Information Management Ontology (PIMO). Proper recommendations, such as recommendations for additional items to add to the pile, apparently should be based on the textual content of the items in the pile.

Although methods of natural language processing for information retrieval could be useful, the most important type of textual processing are those which allow us to relate concepts in PIMO to the processed texts. Since any given PIMO will change over time, this type of natural language processing cannot be performed as preprocessing of all textual context related to the user. Hybrid recommendation needs on-the-fly textual processing with the ability to aggregate the current instantiation of PIMO with the results of textual processing.

Modeling this ontology as a multidimensional network allows the augmentation of the ontology with new information, such as the "semantic" content of the textual information in user documents. Recommendations in Nepomuk-Simple are computed on the fly by graph-based methods performing in the unified multidimensional network of concepts from PIMO augmented with concepts extracted from the documents pertaining to the activity in question.

In [Troussov et al., 2008] Nepomuk-Simple recommendations were classified into two major types. The first type of recommendations is the recommendation of additional items to the pile, when the user is working on an activity. The second type of recommendation arises, for instance, when the user is browsing the Web: Nepomuk-Simple can recommend that the current resource might be relevant to one or more activities performed by the user.

Application for Dynamic Semantic Publishing

The Internet is changing how news is being consumed, and news organizations need to respond with dynamic, responsive, timely, relevant, and quality online content. Content providers want to be able to dynamically compose new web pages in response to external events, to ensure semantic interoperability of their content, to improve navigation, to facilitate content re-use and re-purposing, to reduce overall publishing costs, and to introduce new publishing models at zero incremental effort to the journalists.

"BBC News, BBC Sport and a large number of other web sites across the BBC are authored and published using an in-house bespoke content management/production system ("CPS") with an associated static publishing delivery chain. …The first significant move away from the CPS static publishing model by the BBC's Future Media department was through the creation of the BBC Sport World Cup 2010 website" [Rayfield, 2012].

The BBC's FIFA World Cup 2010 website ([BBC World Cup 2010]) has more than 700 hundreds of topically composed pages for individual football entities such as football teams, groups and players aggregated via semantic web technologies. BBC on-line team coined a new technical term to describe this technology to automate the aggregation and publication of interrelated content objects - "Dynamic Semantic Publishing" [Nowack, 2010].

The usual bottleneck for proliferation of semantic web is the need for manual and tedious work for annotation. IBM LanguageWare, which is now part of IBM Content Analytics, is the text analysis technology that is being used in this project by BBC on-line team to overcome this bottleneck ([MacManus, 2012]).

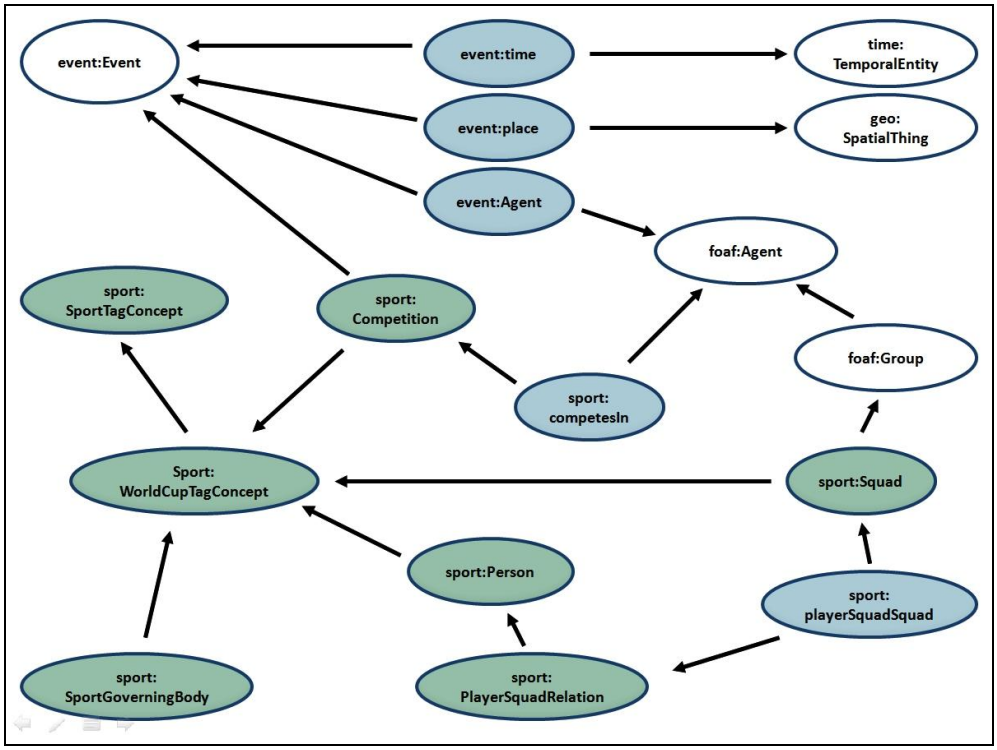

Fig. 4. Domain ontology used in the BBC FIFA World Cup 2010 Website [BBC World Cup 2010], simplified for brevity ([Rayfield, 2012])

The BBCs Fluid Operations' Information Workbench supports the editorial process for the BBC's Dynamic Semantic Publishing strategy, from authoring and curation to publishing of ontology and instance data following an editorial workflow [Zaino, 2012]. According to Peter Haase, Lead Architect R&D at Fluid Operations, this Information Workbench, as deployed by the BBC, "integrates and interlinks dynamic and semantically enriched

data in a central place. Approved content is then available for automatic publication on the website. The platform seamlessly integrates into already existing editorial processes and automates the creation and delivery of semantically enriched content." [Zaino, 2012]

The basis of this system was an ontology that described how World Cup facts related to each other. For example, "Frank Lampard" was part of the "England Squad" and the "England Squad" competed in "Group C" of the "FIFA World Cup 2010". The ontology also included "journalist-authored assets" such as stories, blogs, profiles, images, video and statistics.

IBM LanguageWare provides natural language processing of all related documents w.r.t. this knowledge to identify key concepts and provide named entities disambiguation. IBM Content Analytics analyzes news, as it is created by journalists, and identifies key concepts (explicit or inferred) to a high level of accuracy. "When a journalist writes a story, an athlete is surfaced in suggestions to tag – when someone no one ever heard of before wins a gold medal, he is immediately identified. " (Jem Rayfield, Lead Technical Architect for the News and Knowledge Core Engineering department of BBC, [Zaino, 2012]).

Conclusions and Future Work

We described the architecture and algorithms of ontology based semantic processing used by IBM LanguageWare in several projects and outlined two applications of this technology: one for individual information consumer (Nepomuk-Simple semantic desktop) and another for the information centric organisation (the BBC's World Cup 2010 website). This reusable approach could be also useful in other applications. For instance, in [Kinsella et al., 2008] it was used for social network analysis application such as navigation in the ego-centric networks. Real life applications show that LanguageWare implementation is efficient and scalable; the paper [Judge et al., 2007] described experimental results for graph-based part of the system: network flow operations needed to process a text using semantic networks with several hundred thousand concepts take 200msc on an ordinary PC.

Future work might require additional elaboration on the dictionary layout, on the dynamic update of lexical layer, and on various issues that arise when applying two-layer approach to languages more morphologically rich and linguistic complex than English (see [Troussov et al., 2009a]). Since the major graph-mining operation in IBM LanguageWare is based on a network flow methods, which is very broad class of algorithms, finding the most suitable algorithms for the specific tasks and the topology of the knowledge network is still a challenge.

Bibliography

[Ou et al, 2005] Ou, Weiqiang, Elsayed, Adel, and Hartley, Roger. Towards ontology-based semantic processing for multimodal active presentation. Games Computing and Creative Technologies: Conference Papers, 2005.

[Gruber, 1993] Gruber, T.R. A translation approach to portable ontologies. Knowledge Acquisition, 5(2):199-220, 1993.

[Guarino, 1998] Guarino N. Formal Ontology in Information Systems. In Proceedings of FOIS'98, Trento, Italy, 6-8 June 1998. Amsterdam, IOS Press. 3-15.

[Brin and Page, 1998] Brin, S. and Page, L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: Seventh International World-Wide Web Conference (WWW 1998), April 14-18, 1998, Brisbane, Australia.

[Langville and Meyer, 2006] Langville, A.N. and Meyer, C. Google's PageRank and Beyond: The Science of Search Engine Rankings. Princeton University Press, 2006

[Troussov et al., 2011a] Troussov, A., Darena, F., Zizka, J., Parra, D., and Brusilovsky, P. Vectorised Spreading Activation Algorithm for Centrality Measurement. Acta univ. agric. et silvic. Mendel. Brun. Brno, 2011

[Borgatti, 2005] Borgatti, S. P. Centrality and network flow. Social Networks, 27, 2005, 1: 55–71.

[Borgatti and Everett, 2006] Borgatti, S. and Everett, M. A graph-theoretic perspective on centrality. Social Networks, 28(4):466–484, 2006.

[Troussov et al., 2009] Troussov, A., Levner. E., Bogdan, C., Judge, J., and Botvich, D. Spreading Activation Methods. In Shawkat A., Xiang, Y. (eds). Dynamic and Advanced Data Mining for Progressing Technological Development, IGI Global, USA, 2009.

[Judge et al, 2008] Judge, J., Nakayama, A., Sogrin, M., Troussov, A. Method and System for Finding a Focus of a Document. United States Patent 7870141, Filing Date: 02/26/2008.

[Troussov et al., 2008] Troussov, A., Judge, J., Sogrin, M., Bogdan, C., Lannero, P., Edlund, H., Sundblad, Y. Navigation Networked Data using Polycentric Fuzzy Queries and the Pile UI Metaphor. Proceedings of the International SoNet Workshop (2008), pp. 5-12.

[Rayfield, 2012] Rayfield, J. Sports Refresh: Dynamic Semantic Publishing. Retrieved June 30, 2012, from http://www.bbc.co.uk/blogs/bbcinternet/2012/04/sports_dynamic_semantic.html

[BBC World Cup 2010] BBC World Cup 2010 Website. Retrieved June 30, 2012, from http://news.bbc.co.uk/sport2/hi/football/world_cup_2010/default.stm

[Nowack, 2010] Nowack, B. Dynamic Semantic Publishing for any Blog (Part 1). 2010. Retrieved June 30, 2012, from http://bnode.org/blog/2010/07/30/dynamic-semantic-publishing-for-any-blog-part-1

[MacManus, 2012] MacManus, R. BBC World Cup Website Showcases Semantic Technologies. July 2010. Retrieved June 30, 2012, from http://www.readwriteweb.com/archives/bbc_world_cup_website_semantic_technology.php

[Zaino, 2012] Zaino, J. Sports are the Semantic Focus in Britain at the BBC and in Brazil at Globo. 2012. Retrieved June 30, 2012, from http://semanticweb.com/sports-are-the-semantic-focus-in-britain-at-the-bbc-and-in-brazil-at-globo_b29040

[Kinsella et al., 2008] Kinsella, S., Harth, A., Troussov, A., Sogrin, M., Judge, J., Hayes, C., and Breslin, J.G. Navigating and Annotating Semantically-Enabled Networks of People and Associated Objects. Why Context Matters: Applications of Social Network Analysis (T. Friemel, ed.), VS Verlag, 2008, ISBN 3531163280, 79-96.

[Judge et al., 2007] Judge, J., Sogrin, M., and Troussov, A. Galaxy: IBM Ontological Network Miner. Proceedings of the 1st Conference on Social Semantic Web (CSSW), September 26-28, 2007, Leipzig, Germany.

[Troussov et al., 2009a] Troussov, A., Judge, J., Sogrin, M., Akrout, A., Davis, B., and Handschuh, S. A Linguistic Light Approach to Multilingualism in Lexical Layers for Ontologies. SLT, vol 12, Polish Phonetics Association, ed. G. Demanko, K. Jassem, S. Szpakowicz

Authors' Information

Alexander Troussov *– Ph.D., IBM Dublin Center for Advanced Studies Chief Scientist. Dublin Software Lab, Building 6, IBM Technology Campus, Damastown Ind. Est., Mulhuddart, Dublin 15, Ireland; e-mail: arousso@ie.ibm.com*

Major Fields of Scientific Research: natural language processing, software technologies, network analysis

D.J. McCloskey *– NLP Architect, IBM Watson. Building 6, IBM Technology Campus, Damastown Ind. Est., Mulhuddart, Dublin 15, Ireland; e-mail: dj_mccloskey@ie.ibm.com*

Major Fields of Scientific Research: computational linguistics, natural language processing, semantic web applications