Department of
Information Engineering
and Computer Science
DISI

UNIVERSITY
OF TRENTO - Italy

DISI - Via Sommarive 14 - 38123 Povo - Trento (Italy)
http://www.disi.unitn.it

LIVING KNOWLEDGE

Fausto Giunchiglia, Vincenzo Maltese,
Anthony Baldry, Biswanath Dutta

March 2012

Technical Report # DISI-12-015

Also in: "Web genres and web tools: contributions from the Living Knowledge project", edited by Giulia Magazzù et al., IBIS, Como-Pavia.

Living Knowledge

Fausto Giunchiglia [1], Vincenzo Maltese [1], Anthony Baldry [2], Biswanath Dutta [1]
1 University of Trento; 2 University of Pavia/Messina

1. Introduction

Diversity, especially manifested in language and knowledge, is a function of local goals, needs, competences, beliefs, culture, opinions and personal experience. The LivingKnowledge project considers diversity as an asset rather than a problem. With the project, foundational ideas emerged from the synergic contribution of different disciplines, methodologies (with which many partners were previously unfamiliar) and technologies flowed in concrete **diversity-aware applications** such as the **Future Predictor** and the **Media Content Analyser** providing users with better structured information while coping with Web scale complexities. The key notions of **diversity, fact, opinion** and **bias** have been defined in relation to three methodologies: Media Content Analysis (MCA) which operates from a social sciences perspective; Multimodal Genre Analysis (MGA) which operates from a semiotic perspective and Facet Analysis (FA) which operates from a knowledge representation and organization perspective. A **conceptual architecture** that pulls all of them together has become the core of the tools for automatic extraction and the way they interact. In particular, the conceptual architecture has been implemented with the Media Content Analyser application. The scientific and technological results obtained are described in the following.

2. Innovative ways to manage diversity in knowledge

The world is extremely diverse and diversity is visibly manifested in language, data and knowledge. The same real world object can be referred to with many different words in different communities and in different languages. For instance, it is widely known that in some Nordic circumpolar groups of people the notion of snow is denoted with hundreds of different words in the local language carrying very fine grained distinctions [1]. This phenomenon is often a function of the role and importance of the real world object in the life of a community. Conversely, the same word may denote different notions in different domains; e.g. bug as insect in entomology and bug as a failure or defect in a computer program in computer science. Space, time, individual goals, needs, competences, beliefs, culture, opinions and personal experience also play an important role in characterizing the meaning of a word. Diversity is an unavoidable and intrinsic property of the world and as such it cannot be avoided. At the same time, diversity is a local maximum since it aims at minimizing the effort and maximizing the gain [2].

To make diversity traceable, understandable and exploitable, to favour interoperability and let people as well as machines *understand* it is therefore essential to provide effective ways to make the meaning of the words in a certain context (i.e. their semantics) explicit, such that information becomes unambiguous to humans as well as to machines. Towards this goal we believe that a preliminary step is

the creation of a **diversity-aware knowledge base**. This requires appropriate methodologies for its representation, construction and maintenance. A knowledge base can be seen a collection of facts encoding knowledge of the real world that can be used to automate tasks. Nevertheless, to be useful, a knowledge base should be very large, virtually unbound and able to capture the diversity of the world as well as to reduce the complexity of reasoning at run-time. At this purpose, the notions of **domain** (as originated from library science) and **context** (as originated from Artificial Intelligence) have been indicated as essential for diversity-aware knowledge bases [3].

Domains have two important properties. They are the main means by which diversity is captured, in terms of language, knowledge and personal experience. For instance, according to the personal perception and purpose, the *space* domain may or may not include buildings and man-made structures; the *food* domain may or may not include dogs according to the local customs. Moreover, domains allow scaling as with them it is possible to add new knowledge at any time as needed. For instance, while initially local applications may require only knowledge of the *space* domain, due to new scenarios, the *food* domain might be needed and added.

Determining the context allows on the one hand a better disambiguation of the terms used (i.e. by making explicit some of the assumptions left implicit) and on the other hand, by selecting from the domains the language and the knowledge which are strictly necessary to solve the problem, it allows reducing the complexity of reasoning at run-time.

Diversity was also formalized in terms of **diversity dimensions**, i.e. the dimensions by which knowledge is framed. In library science *topic*, *space* and *time* are known to be the three fundamental diversity dimensions [4]. We adopted this view also in the project and started providing support for them. This has been done in particular within GeoWordNet [14] and T-YAGO [21, 71].

Only a few existing knowledge bases can be considered diversity-aware. In Table 1 we summarize the characteristics of existing knowledge bases. The table also includes Entitypedia, the diversity-aware knowledge base that we have been developing centered on these two important notions. It is completely modular since at any moment it allows plugging an arbitrary number of domains; it has a clear split between classes (concepts), entities (their instances), their relations and attributes; it provides different vocabularies in different languages (initially in English and Italian) that are neatly distinguished from the formal language used in task automation. A first version of Entitypedia was presented at the ICT2010 expo in Bruxelles.

For the construction of the domain knowledge a new methodology called **DERA** was proposed. The DERA framework is entity oriented and it is meant to develop domains to be used for automation. It is being developed bearing real world entity representations in mind, including inter-alia locations, people, organizations, songs, movies, relevant to a given domain. It is based and adapted from the faceted approach [4], an effective methodology for domain construction and maintenance. In fact, as it is proved by decades of research in library science, the use of the principles at the basis of the faceted approach guarantees the creation of better quality - in terms of robustness, extensibility, reusability, compactness and flexibility - and easier to maintain domain ontologies [11, 12, 13]. We also developed a user

interface for the creation and maintenance of domains as a support for the DERA methodology. As described in [3], by using the DERA methodology Entitypedia has been incrementally populating with domain knowledge starting with *space*. By taking GeoNames, WordNet and MultiWordNet as main sources, the work on *space* led to the creation of GeoWordNet [14, 15] a very large open source geo-spatial ontology containing overall more than 1000 classes, 7 million entities, 70 different kinds of relations and 35 kinds of attributes. The facets built from classes include region, administrative division, populated place, facility, abandoned facility, land, landform, body of water, agricultural land and wetland. However, the long term goal is not to build the whole world knowledge, but to identify those domains which are more likely to play a role in everyday life and in particular on the Web. This has been identified as strategic towards enabling diversity-aware applications for the Web. A prioritized list of around 350 domains was identified. On the very top of this list we find domains such as *space*, *food*, *sport*, *tourism*, *music*, *movie* and *software*. They were called **Internet domains** or also **everyday domains**.

Knowledge base	#entities	#facts	Domains	Distinction concepts instances	Distinction natural language formal language	Manually built
YAGO [5]	2.5 M	20 M	No	No	No	No
CYC [6]	250,000	2.2 M	Yes	No	No	Yes
OpenCYC [6]	47,000	306,000	Yes	No	No	Yes
SUMO [7]	1,000	4,000	No	Yes	Yes	Yes
MILO [8]	21,000	74,000	Yes	Yes	Yes	Yes
DBPedia [9]	3.5 M	500 M	No	No	No	No
Freebase [10]	22 M	unknown	Yes	Yes	No	Yes
Entitypedia [3]	10 M	80 M	Yes	Yes	Yes	Yes

Table 1: Overview of existing knowledge bases

By incrementally providing knowledge in multiple domains at three different levels - natural language, formal language, knowledge - and by following a precise methodology in domain construction and maintenance, Entitypedia offers an effective support to diversity. Initial experiments in this direction include:

- The *political science* domain, developed to annotate a corpus of documents and used by the faceted search facility within the Media Content Analyser application [16]
- The collaboration with Telecom Italia for the analysis of their *food* domain using DERA principles, which shows how they allow identifying typical pitfalls and mistakes and giving concrete suggestions to improve the ontology

- The usage of a customized version of the *space* domain in the geo-catalogue of the Province of Trento in Italy, where the evaluation shows that with the use of the faceted ontology and the semantic matching open source tool S-Match [19] it is possible to more than double the number of correct documents found by the search facility [17, 18]

One drawback of the approach followed with Entitypedia is that, in order to guarantee the high quality of the knowledge, its construction and maintenance requires a significant amount of manual work. In fact, building a domain may take several weeks of work by an expert familiar with the DERA methodology. For instance, bootstrapping the *space* domain (that, given its pervasiveness, is among the biggest ones) took around 6 man months. However, other domains should take much less. We plan to overcome this issue by adopting crowdsourcing techniques integrated with a certification pipeline based on ideas already exploited on ESP games [20]. Given the precise split enforced between concepts and instances, the plan is to establish two pipelines: the first for experts at the purpose of defining the basic terminology of domains, in terms of classes, relations and qualities (the TBox); the second for generic users at the purpose of providing actual data for the entities (the ABox). The main reason for this distinction is that the first requires a higher level of expertise. At this purpose, some training activities have been already conducted at the ISI Institute in India where some students in library science were asked to use the DERA methodology for the construction of sample domains. Notice how the second pipeline will have to be able to manage a quantity of knowledge which is several orders of magnitude bigger that the first. When possible, given format and quality of the data, ready-made entities can be directly imported from existing sources. This is for instance what has been done for the population of the space domain where, after the manual analysis of the classes, the entities were automatically imported from GeoNames [14].

3. Characterizing and handling bias

Two definitions of bias are typically found in contemporary dictionaries of English [22]. The first, "*Bias is prejudice against one group and favouritism towards another, which may badly affect someone's judgement of a situation or issue",* emphasizes the tight relationship between bias and social diversity as expressed in such two-word combinations as *gender bias*, *racial bias*, *media bias* and where bias is, in part, synonymous with prejudice and partiality. The second, "*Bias is a concern with or interest in one thing more than others",* is reflected in such expressions as *interviewer bias, follow-up bias, misclassification bias.* While the first definition reflects **distortions in the *quality* of information**, the second describes **distortions in the *quantity* of information**. Thus, the second type of bias has to do mainly with referencing and quantifying suppression or restrictions in information arising from preferences, decisions and procedures, often in scientific or semi-scientific research.

Given the way in which society works, the two meanings are far from being mutually exclusive. The degree of linkage and interdependence arises mainly from the effects of two other elements in the 'equation': (a) the type or 'mode' of information involved**,** in particular whether written, spoken, visual

or multimedia combinations of these, and (b) the effects of context. We showed how four parameters are relevant to the detection of bias:

- **quality of information** with its implicit ties with diversity;
- **quantity of information** with its implicit ties with deceitful or accidental suppression;
- **type/mode of information** with its implicit ties with the resources and media used;
- **context** with its implicit capacity to prioritize one of the previous parameters over others.

With this in mind, the traditional distinction between the two traditional definitions of bias is maintained, so that our definition of bias transcends the characterization of one particular type of socially-recognized bias as opposed to others. Thus, for example, the definition is not restricted to media bias, despite the fact that its popularization of political and social conflict and dissent has led to its prominence in many research studies [23]. Instead, it unpacks and systematizes these relationships mainly in relation to the notion of bias in semiotics. The unitary definition of bias we gave is:

Bias arises in relation to the quality or quantity of information (or both). As regards the former, bias is the degree of correlation between (a) the polarity of an opinion and (b) the context of the opinion holder. As regards the latter, bias is a failure to provide sufficient information for the purposes of establishing the context of the opinion holder at one of the relevant levels. The context of the opinion-holder is potentially defined concurrently at 4 levels: the text level (word, sentence etc.); the co-text level, the context of situation level and the context of culture level. The actual contribution of each level to the definition of bias varies from instance to instance.

Context is a decisive parameter. Different types of context are deemed to exist in semiotics and/or linguistics interpretable as a series of levels, each of which provides a wider and clearer view of events, actions and activities on which to base analytical judgments. They are the *sentence*, *co-text* [22, 28, 29], *context of situation* and *context of culture* [24, 25, 26, 27]. On the other hand, the *minimum amount* of context needed to establish the opinion holder's context in specific texts is a desirable methodological objective; it is in keeping with the principle of economy and efficiency in scientific interpretation. Bias as distortion in the quantity of information arises when one or more of the levels of context above have been suppressed.

Bias is highly resistant to scientific measurement both from a qualitative and quantitative standpoint as it is concerned with the opinion-based end of the fact-opinion cline. Nevertheless, drawing on the interdisciplinary forces the definition of bias given has also made it possible to define self-regulatory parameters guiding the systematic treatment of bias. The definition of bias within the framework of discourse context provides a way of negotiating bias in terms of part/whole relationships in web pages. The focus can thus be on **micro strategies** (such as MGA: mini-genre analysis) which explore specific recurrent ***parts*** of web pages (logos, photos, captions); or it can be on **macro-strategies** where the focus is on ***whole*** page analysis which explores user/reader responses to bias on specific issues (such as their detection of writers' prejudices in relation to immigration/integration, climate change etc.); it can also be, and ultimately has been, on the integration of both strategies. The complementarity of **micro** and **macro** approaches to bias handling and bias detection ensues from the fact that bias arises, not from a

single source, but from the interplay of different types/levels of interacting discourse units/contexts found on any web page.

The case studies we have explored contain abundant examples of both factual and opinion-holder/identity omission. They are expected to help in the process of exemplifying how the technology can be put to use.

A first example relates to the fact that bias by omission lends itself to clustering in relation to topic and time so that scanning for omission of high-value facts in the cluster is possible (e.g. in relation to future prediction). A specific example of deliberate omission relates to the possibility of extracting fact suppression from a cluster (temporal, entity and topic clustering) of, say, 15 articles from one week in the trial of a corrupt politician favoring a company; the software might well highlight the fact that, unlike the others, two articles omit both that he was formerly a director of the company and still receives money from it and that his daughter is on the company board. This approach would seem to be applicable to the use cases developed within the project (e.g. immigration, climate change, Nabucco pipeline) where newspaper reports exemplifying this type of bias are not uncommon.

A second example based on work already carried out within the project relates more specifically to the differences in British quality newspaper and tabloid reporting, for example, the arrest of the would-be assassin of a Danish cartoonist on 01/01/2010 in which fairly traditional discourse analysis shows omission of facts about the participants (police/cartoonist/assailant) according to each newspaper's political bias and expectations about readers' literacy skills and political/cultural views (e.g. Islam). In this respect, the capacity of LivingKnowledge to carry out detection in relation to bias extends beyond the linguistic and includes the visual. Visual analysis of the same newspaper reporting of the assassination attempt, for example, shows that the number, distribution, size and positioning of 'supporting' photographs contributes to the overall patterns of bias found in the linguistic text. Given that the relationship between visual and verbal in terms of bias detection is at the forefront in the project, further progress in this field can be expected. For example, work on logo detection shows that logos, as mainly visual genres, work to make their meaning at the context of culture level, i.e. establishing a particular company as being a paragon of virtue in the minds of people in many countries. However, companies and/or political parties are likely to change or suppress the use of an existing logo in cases where press reporting is unfavorable replacing it with other textual devices, not necessarily a replacement logo or even a visual device.

4. Developed tools

The diversity-aware tools developed with LivingKnowledge have been deployed in a **testbed** which provides a scalable technology for feature extraction, annotation, indexing, search and navigation on very large multimodal datasets. The testbed also includes a huge collection of web documents of around 10TB of raw data and 100GB of processed data. The testbed is an open source project (see www.diversityengine.org) and is available for experimentation, dissemination and exploitation [60]. Specific research was conducted to support the functionalities of the testbed. In particular, it includes

efficient mechanisms for search, both on structured data (retrieving and ranking over structured data [62], repeatable and reliable search system evaluation using crowdsourcing [61], co-reference aware object retrieval [63], keyword search over RDF graphs [64]) and on large collections and web (timestamp-based result caching [58] or adaptive Time-to-Live strategies for query result caching [59]), as well as efficient indexing [57].

Basic fact and opinion extraction from text and images. Since the Web is multimodal in nature, work has been done on both text and image processing. An initial set of tools was integrated for text analysis, including traditional natural language processing tasks such as part of speech (POS) tagging, named entity recognition (NER) and syntactic parsing. Facts were extracted in a way to position them in space and time [21] and advances were made in named entities disambiguation [38, 43]. The work on opinions aimed at *polarity* estimation and classification (positive/negative/neutral) at sub-sentence, sentence and document level [34, 35] and *opinion holder* detection [39, 40]. As underlined in the previous section, this is of fundamental importance towards the detection of bias. Concerning images, a significant amount of work has been devoted to the usage of image forensic techniques to determine if images were tampered in some way. It includes copy-move forgery detection [30, 31, 33, 41, 42], photomontage detection [32], e.g. of faces, and image reuse analysis. Overall, more than 50 analysis tools (for text and images) and visualization components have been developed.

Diversity-aware tools. Several tools able to analyze and place documents along diversity dimensions were developed [44, 45]. Most of the work focused on the three fundamental dimensions: topic, space and time. Significant work has been done towards *diversification of search results* [47, 48] and in particular on methods about topical diversification [46], sentiment analysis for controversial topics [49, 72, 73, 74] and *incremental diversification* of very large sets based on data streams [55]. They are often based on techniques to efficiently detect semantically near-duplicate documents. Some of these tools, methods and corresponding experimental results (for both text and images) are described in [50, 51, 52, 53, 54]. Some early work on bias detection focused on methods to compare parliament speeches with news articles [56]. Diversification along the temporal dimension was done by extracting facts together with their temporal information when possible. T-YAGO [71] is an extension of the YAGO ontology [70] where facts are placed in space and time. The LK-Finder tool [38, 43, 21] was designed to extract and disambiguate temporal expressions and related named entities from text. CATE [75, 76] allows visualizing them on a timeline. Tools to determine how images are reused over time by different documents within a corpus have been also experimented [36, 37]. Applications and methods to cluster, classify and aggregate data by diversity dimension were also developed [77, 80]. Taking advantage from a variety of feedback mechanisms based on annotating, commenting and rating of content, tools were developed for prediction of photo attractiveness [82], privacy and sentiment [78], prediction of comment ratings for videos [79], personalized suggestions [81, 84] and aggregation of sentiment information [83]. Work on indexing, matching and clustering images (mainly based on automatic geo-location identification) was based on SIFT features [36] and won the Open Source Software Challenge at ACM Multimedia 2011 [37].

Diversity-aware applications. The testbed is at the basis of the two applications we developed: the Future Predictor and the Media Content Analyser.

The Future Predictor [67], or Time Explorer, allows searching for statements about the past, present or future. Results are visualized on a timeline and can be navigated by different facets [65, 66], e.g. by topic or named entity (e.g. person, organization, event). Specific research in this field enhanced search and retrieval technology [68, 69] by providing information aggregation and summarization, switching from document search to retrieval of factual knowledge, providing factual answers together with supporting documents, enabling searches with reference to the past, present or future, clustering of search results based on diversity/viewpoints, ranking of search results not only on popularity, but also on diversity.

The Media Content Analyser is a valid tool to assist coders towards conducting typical MCA studies. Typical questions that can be addressed by the tool, e.g. on integration, include:

- What are the main [*concepts, people, parties, countries, dates, resolutions, etc.*] related to integration?
- Which of these [*concepts, people, parties, countries, dates, resolutions, etc.*] are most [*controversial, accepted, subjective, biased, etc.*]?
- What are the main parties discussing integration in a *negative, positive, controversial, etc.*] context?
- Which parties have changed their discourse on integration (i.e. from positive to negative)?
- Which politicians have changed their discourse on integration (i.e. from positive to negative)?
- Which periods of time are most important vis-à-vis integration, and how are other events correlated to these periods?

With the Media Content Analyser application many automatic feature extraction algorithms, available through the testbed, can be applied to documents in order to speed up the annotation of text and images (previously only manually performed). They include named entity recognition and disambiguation, statement and speaker extraction, image extraction and face detection. The corpus of documents can be navigated as a list, as a timeline or by facet.

The Future Predictor and the Media Content Analyser represent a clear step forward in the research on how to understand and detect bias and diversity taking evolution of knowledge into account for retrieval, and allowing opinion and diversity-aware search.

References

1. Artic Climate Impact Assessment, Cambridge University Press, 2005, pp. 973.
2. Giunchiglia, F. (2006). Managing Diversity in Knowledge, Invited Talk at the European Conference on Artificial Intelligence ECAI, Lecture Notes in Artificial Intelligence 2006.
3. Giunchiglia, F., Maltese, V., Dutta, B. (2012). Domains and context: first steps towards managing diversity in knowledge. Journal of Web Semantics, special issue on Reasoning with Context in the Semantic Web. DOI: 10.1016/j.websem.2011.11.007

4. Ranganathan, S. R. (1967). Prolegomena to library classification, Asia Publishing House.
5. F. M. Suchanek, G. Kasneci, G. Weikum, YAGO: A Large Ontology from Wikipedia and WordNet, Journal of Web Semantics (2011).
6. R. Guha, D. Lenat, Context dependence of representations in cyc, Colloque ICO (1993).
7. C. Matuszek, J. Cabral, M. Witbrock, J. DeOliveira, An introduction to the syntax and content of Cyc, AAAI Spring Symposium (2006).
8. A. Pease, G. Sutcliffe, N. Siegel, S. Trac, Large theory reasoning with SUMO at CASC, AI Communications, 23 2-3 (2010) 137–144.
9. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, C. Cyganiak, Z. Ives, DBpedia: A Nucleus for a Web of Open Data, 6th International Semantic Web Conference ISWC (2007).
10. K. Bollacker, C. Evans, P. Paritosh, T. Sturge, J. Taylor, Freebase: a collaboratively created graph database for structuring human knowledge, ACM SIGMOD international conference on Management of data (2008), 1247-1250.
11. V. Broughton, The need for a faceted classification as the basis of all methods of information retrieval, Aslib Proceedings 58 1/2 (2006), 49-72.
12. V. Broughton, Building a Faceted Classification for the Humanities: Principles and Procedures, Journal of Documentation (2007).
13. L. Spiteri, A Simplified Model for Facet Analysis, Journal of Information and Library Science 23 (1998), 1-30.
14. F. Giunchiglia, V. Maltese, F. Farazi, B. Dutta, GeoWordNet: a resource for geo-spatial applications, Extended Semantic Web Conference ESWC (2010).
15. B. Dutta, F. Giunchiglia, V. Maltese, A facet-based methodology for geo-spatial modelling, GEOS (2011).
16. D. P. Madalli, A.R.D. Prasad, Analytico synthetic approach for handling knowledge diversity in media content analysis, UDC seminar (2011).
17. F. Farazi, V. Maltese, F. Giunchiglia, A. Ivanyukovich. A faceted ontology for a semantic geo-catalogue. In Proceedings of the ESWC 2010.
18. F. Farazi, V. Maltese, F. Giunchiglia, A. Ivanyukovich. Extending a geo-catalogue with matching capabilities. At the LHD Workshop at IJCAI 2011.
19. Giunchiglia F, Autayeu A, Pane J (2010) S-Match: an open source framework for matching lightweight ontologies. The Semantic Web journal, 2010.
20. L. Von Ahn. Games With A Purpose, IEEE Computer Magazine (2006), 96-98.
21. J. Hoffart, F.M. Suchanek, K. Berberich, E. Lewis-Kelham, G. de Melo, and G. Weikum, YAGO2: exploring and querying world knowledge in time, space, context, and many languages. In Proceedings of the 20th international conference on World Wide Web 2011, Hyderabad, India.
22. J. Sinclair (ed.) (2001). Collins COBUILD English Dictionary for Advanced Learners. Third edition. Glasgow: HarperCollins Publishers.
23. Anand, Bharat, Di Tella, Rafael, Galetovic, Alexander, 2007. Information or Opinion? Media Bias as Product Differentiation. Journal of Economics & Management Strategy, Volume 16, Number 3, Fall 2007, 635–682.
24. M. Halliday (1961). Categories of the theory of grammar. Word 17, 3: 241-92.
25. M. Halliday (1978). Language as a Social Semiotic. London: Arnold.
26. M. Halliday, R. Hasan (1998). Language, Context, and Text: Aspects of Language in a Social-Semiotic Perspective (Language Education). Oxford: OUP.
27. M. Halliday, C. Matthiessen (2004). An Introduction to Functional Grammar (3rd ed.). London: Arnold.
28. J. Sinclair (1991). Corpus, Concordance, Collocation. Oxford: Oxford University Press.

29. J. Sinclair (2008). The phrase, the whole phrase and nothing but the phrase. In Sylviane Granger and Fanny Meunier (eds), Phraseology: An interdisciplinary perspective. Amsterdam: John Benjamins, pp. 407-10.
30. T. Bianchi, A. De Rosa, A. Piva, Improved DCT coefficient analysis for forgery localization in JPEG images, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, May 22 - 27, 2011, Prague, Czech Republic, pp. 2444-2447
31. T. Bianchi, A. Piva, Detection Of Non-Aligned Double Jpeg Compression with Estimation Of Primary Compression Parameters, IEEE International Conference on Image Processing, ICIP, Sept 11-14, 2011, Brussels, Belgium
32. V. Conotter, G. Boato, Analysis of sensor fingerprint for source camera identification, Electronics Letters, v. 47, n. 25, 2011.
33. M. Fontani, T. Bianchi, A. De Rosa, A. Piva, M. Barni, A Dempster-Shafer Framework for Decision Fusion in Image Forensics, IEEE Intl. Workshop on Information Forensics and Security (WIFS'11), Foz do Iguaçu, Brazil, November 29th - December 2nd, 2011
34. Sucheta Ghosh et al., Shallow Discourse Parsing with Conditional Random Fields, Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP). Chiang Mai, Thailand, 2011.
35. Sucheta Ghosh et al, End-to-End Discourse Parser Evaluation, Proceedings of the Fifth IEEE International Conference on Semantic Computing (ICSC 2011). Palo Alto, United States, 2011.
36. Hare, J., Samangooei, S. Lewis, P, 2011. Efficient clustering and quantisation of SIFT features: Exploiting characteristics of the SIFT descriptor and interest region detectors under image inversion, The ACM International Conference on Multimedia Retrieval (ICMR 2011), 17-20 April 2011, Trento, Italy
37. Hare, J., Samangooei, S. and Dupplaw, D. “OpenIMAJ and ImageTerrier: Java Libraries and Tools for Scalable Multimedia Analysis and Indexing of Images”. In: ACM Multimedia 2011, 28/11/2011 until 1/12/2011, Scottsdale, Arizona, USA. pp. 691-694. 2.
38. J. Hoffart, M. A. Yosef, I. Bordino, H. Fuerstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, G. Weikum. Robust disambiguation of named entities in text. In EMNLP 2011
39. Richard Johansson and Alessandro Moschitti, Syntactic and Semantic Structure for Opinion Expression Detection. In Proceedings of the 2010 Conference on Natural Language Learning, Upsala, Sweden, July 2010. Association for Computational Linguistics.
40. Richard Johansson and Alessandro Moschitti, Extracting opinion expressions and their polarities – exploration of pipelines and joint models. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 101–106, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.
41. O.Muratov, P. Zontone, G. Boato, F. G. B. De Natale, “A Segment-based Image Saliency Detection”, International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011
42. Oleg Muratov, Duc-Tien Dang-Nguyen, Giulia Boato, Francesco G.B. De Natale, Saliency Detection as a Support for Image Forensics, IEEE International Symposium on Communications, Control and Signal processing 2012
43. M. A. Yosef, J. Hoffart, I. Bordino, M. Spaniol, G. Weikum: AIDA: An Online Tool for Accurate Disambiguation of Named Entities in Text and Tables. In PVLDB 4(12): 1450-1453 (2011)
44. K. Denecke, 2009. Assessing content diversity in medical weblogs. First International Workshop on Living Web: Making Web Diversity a true asset held in c onjunction with the International Semantic Web Conference.
45. V. Maltese, F. Giunchiglia, K. Denecke, P. Lewis, C. Wallner, A. Baldry, D. Madalli, 2009. On the interdisciplinary foundations of diversity. First International Workshop on Living Web: Making Web Diversity a true asset held in conjunction with the International Semantic Web Conference.

46. E. Minack, G. Demartini, W. Nejdl, 2009. Current Approaches to Search Result Diversification. First International Workshop on Living Web: Making Web Diversity a true asset held in conjunction with the International Semantic Web Conference.
47. D. Skoutas, M. Alrifai, W. Nejdl, 2009. Re-ranking web service search results under diverse user preferences. In 4th Int. Workshop on Personalized Access, Profile Management, and Context Awareness in Databases.
48. D. Skoutas, E. Minack, W. Nejdl, 2010. Increasing diversity in web search results. In WebSci10: Extending the Frontiers of Society On-Line.
49. G. Demartini, S. Siersdorfer, 2010. Dear search engine: What's your opinion about...? - sentiment analysis for semantic enrichment of web search results. In Semantic Search 2010 (WWW'10 workshop).
50. E. Demidova, P. Fankhauser, X. Zhou, Wolfgang Nejdl, 2010. DivQ: Diversification for keyword search over structured databases. In SIGIR, pages 331–338.
51. E. Ioannou, O. Papapetrou, D. Skoutas, W. Nejdl, 2010. Efficient semantic-aware detection of near duplicate resources. In ESWC, pages 136–150.
52. P. Zontone, M. Carli, G. Boato, F.G.B. De Natale, 2010. Impact of contrast modification on human feeling: an objective and subjective assessment. In ICIP.
53. P. Zontone, G. Boato, F. G. B. De Natale, A. De Rosa, M. Barni, A. Piva, J. Hare, D. Dupplaw, Paul Lewis, 2009. Image diversity analysis: Context, opinion and bias. First International Workshop on Living Web: Making Web Diversity a true asset held in conjunction with the International Semantic Web Conference.
54. A. De Rosa, F. Uccheddu, A. Costanzo, A. Piva, and M. Barni, 2010. Exploring image dependencies: a new challenge in image forensics. In Media Forensics and Security XII Conference, IS&T/SPIE Electronic Imagining.
55. E. Minack, W. Siberski, W. Nejdl, 2011. Incremental Diversification for Very Large Sets: a Streaming-based Approach. In SIGIR.
56. R. Krestel, A. Wall, W. Nejdl, 2012. Treehugger or Petrolhead? Identifying Bias by Comparing Online News Articles with Political Speeches. In: Proceedings of the 21st International Conference on World Wide Web (WWW 2011), Poster, Lyon, France, April 16-20.
57. R. Blanco, E. Bortnikov, F. Junqueira, R. Lempel, L. Telloli, H. Zaragoza. Caching for Incremental Indexes, SIGIR 2010.
58. S. Alici, I. S. Altingövde, R. Ozcan, B. B. Cambazoglu, Ö. Ulusoy. Timestamp-based result cache invalidation for web search engines. In: 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 973-982, Beijing, China, July 2011.
59. S. Alici, I. S.r Altingövde, R.t Ozcan, B. Barla Cambazoglu, Ö. Ulusoy. Adaptive Time-to-Live Strategies for Query Result Caching in Web Search Engines. ECIR 2012.
60. D. Dupplaw, M. Matthews, R. Johansson, Paul Lewis. LivingKnowledge: A Platform and Testbed for Fact and Opinion Extraction from Multimodal Data. Proceedings of the 1st International Workshop on Eternal System (EternalS'11). Budapest, Hungary, 2011.
61. R. Blanco, H. Halpin, D. Herzig, P. Mika, J. Pound, H.Thompson, T. D. Tran. Repeatable and Reliable Search System Evaluation using Crowd-Sourcing. SIGIR, 2011.
62. R. Blanco, H. Halpin, D. M. Herzig, P. Mika, J. Pound, H. S. Thompson, T. Tran Duc. Entity Search Evaluation over Structured Web Data. EOS 2011.
63. J. Dalton, R. Blanco, P. Mika. Coreference Aware Web Object Retrieval. CIKM 2011.
64. R. Blanco, P. Mika, S. Vigna. Effective and Efficient Entity Search in RDF Data. International Semantic Web Conference (1) 83-97 27, 2011.
65. G. Demartini, M. M. S. Missen. R. Blanco, H. Zaragoza. TAER: Time-Aware Entity Retrieval. CIKM 2010.

66. G. Demartini, M. M. S. Missen, R. Blanco. H. Zaragoza. Entity Summarization of News Articles, SIGIR 2010.
67. M. Matthews, J. Atserias, P. Tolchinski, P. Mika, R. Blanco, H. Zaragoza. Searching through time in the New York Times. HCIR Challenge 2010
68. G. Amodeo, R. Blanco, U. Brefeld. Hybrid Models for Future Event Prediction. SIGIR 2011
69. N. Kanhabua, R. Blanco, M. Matthews. Ranking Related News Predictions. SIGIR 2011
70. F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: a core of semantic knowledge. In: WWW 2007
71. Y. Wang, M. Zhu, L. Qu, M. Spaniol, and G. Weikum: Timely YAGO: harvesting, querying, and visualizing temporal knowledge from Wikipedia. EDBT 2010
72. R. Johansson, A. Moschitti. Syntactic and Semantic Structure for Opinion Expression Detection. In CoNLL, 2010.
73. R. Johansson, A. Moschitti. Reranking Models in Fine-grained Opinion Analysis. In COLING 2010
74. R. Awadallah, M. Ramanath, and G. Weikum. Opinionetit: Understanding the opinions-people network for politically controversial topics. In CIKM 2011
75. T. A. Tuan, S. Elbassuoni, N. Preda, and G. Weikum: CATE: context-aware timeline for entity illustration. In WWW 2011
76. A. Mazeika, T. Tylenda, and G. Weikum. Entity timelines: Visual analytics and named entity evolution. In CIKM 2011
77. S. Siersdorfer, J. S. Pedro, M. Sanderson. Content Redundancy in YouTube and its Application to Video Tagging ACM Transactions on Information Systems (TOIS), 2011
78. S. Siersdorfer, J. Hare, E. Minack, F. Deng. Analyzing and Predicting Sentiment of Images on the Social Web (Short Paper) 18th ACM Multimedia Conference (MM 2010), Florence, Italy
79. S. Siersdorfer, J. S. Pedro, S. Chelaru, W. Nejdl. How useful are your comments?- Analyzing and Predicting YouTube Comments and Comment Ratings19th International World Wide Web Conference, WWW 2010, Raleigh, USA
80. S. Siersdorfer, J. S. Pedro, M. Sanderson. Automatic Video Tagging using Content Redundancy 32nd ACM SIGIR Conference, Boston, USA, 2009
81. S. Siersdorfer, S. Sizov. Social Recommender Systems for Web 2.0 Folksonomies 20th ACM Conference on Hypertext and Hypermedia, Hypertext 2009, Torino, Italy
82. S. Siersdorfer, J. S. Pedro. Ranking and Classifying Attractiveness of Photos in Folksonomies 18th International World Wide Web Conference, WWW 2009, Madrid, Spain
83. G. Demartini, S. Siersdorfer, S. Chelaru, W. Nejdl. Analyzing Political Trends in the Blogosphere. Fifth International Conference on Weblogs and Social Media (ICWSM), Barcelona, Catalonia, Spain
84. R. Krestel, P. Fankhauser. Personalized topic-based tag recommendation. Neurocomputing 76(1): 61-70 (2012)