ON THE INTERDISCIPLINARY FOUNDATIONS OF DIVERSITY

Vincenzo Maltese, Fausto Giunchiglia, Kerstin Denecke,
Paul Lewis, Cornelia Wallner, Anthony Baldry, Devika Madalli

July 2009

Technical Report # DISI-09-040

Also: At the First International Workshop on Living Web: Making Web Diversity a true asset (at the ISWC 2009), 25th October 2009, Washington DC, USA

On the interdisciplinary foundations of diversity

Vincenzo Maltese (1), Fausto Giunchiglia (1), Kerstin Denecke (2),
Paul Lewis (3), Cornelia Wallner (4), Anthony Baldry (5), Devika Madalli (6).

(1) DISI Università di Trento, Trento, Italy; (2) L3S Research Center, Hannover, Germany; (3) ECS, University of Southampton, UK; (4) Institute for Social Research and Analysis (SORA), Vienna, Austria; (5) Computer Vision and Multimedia Lab (CVML) University of Pavia, Italy; (6) Indian Statistical Institute, Bangalore, India

Abstract. Designing new-generation diversity-aware tools that face up to the emerging complexity in knowledge is one of the biggest research challenges in recent years. In this paper, we provide key notions about opinion, bias and diversity, and propose an interdisciplinary approach when managing them. Our basic tenet is that diversity should be seen as an asset rather than as a problem to be avoided *a priori*.

Keywords: Knowledge diversity, Knowledge representation and management, Diversity-aware applications

1 Introduction

One of the biggest research challenges in recent years [10] has been facing up to the emerging complexity in data, information and knowledge, in terms of size, diversity of sources, diverging viewpoints, while taking the dynamics of their unpredictable evolution in time into account [1]. The Web is the clearest example of the enormous quantity and diversity of material – text, images and other media – continuously made available online. It is widely agreed that knowledge is strongly influenced by the diversity of context, mainly cultural, in which it is generated [2]. Thus, while it may be appropriate to say that (some kinds of) cats and dogs are food in some parts of China, Japan, Korea, Laos and the Philippines, this is unlikely to be the case in the rest of the world. Sometimes, it is not just a matter of diversity in culture, viewpoints or opinion, but rather a function of different perspectives and goals. In fact, knowledge useful for a certain task, and in a certain environment, will often not be *directly* applicable to other circumstances, and will thus require adaptation. Hence the pressing need to find effective ways of dealing with such complexity, especially in terms of scalability and adaptability in data and knowledge representation. As first advocated in [10], we are firmly convinced that diversity in knowledge should not be avoided, as often happens in approaches where, at design time, a global representation schema is proposed. Rather diversity in knowledge is a key feature, our goal being to develop methods and tools leading to effective design by harnessing, controlling and using the effects of

[1] Details can be found in the ongoing delivery report “Foundations for the representation of diversity, evolution, opinion and bias”. Living Knowledge EU FET project, WP1, 2009

[2] Details can be found in the ongoing delivery report “Analysis of Bias and Diversity: Problems, Features, Related Work”. Living Knowledge EU FET project, WP4, 2009.

emergent knowledge properties. Using these tools, new knowledge can be obtained by adapting existing knowledge but respecting the not entirely predictable process of knowledge evolution and/or aggregation. We envisage a future where developing diversity-aware navigation and search applications will become increasingly important as they will automatically classify and organize opinions and bias producing more insightful, better organized, aggregated and easier-to-understand output by detecting and differentiating between, what we call, diversity dimensions. This explains our adoption of a highly interdisciplinary approach that brings together expertise from a wide range of disciplines: sociology, philosophy of science, cognitive science, library and information science, semiotics and multimodal information theory, mass media research, communication, natural language processing and multimedia data analysis. A solution to the problems posed above is gradually emerging from this synergy.

The rest of the paper is organized as follows. Section 2 explores the key notions of opinion, bias and diversity. Section 3 describes the current state of the art in opinion mining and diversity-aware web searching, while Section 4 describes the proposed framework within the interdisciplinary approach followed. Finally, Section 5 draws some conclusions and outlines future work.

2 Opinion, bias and diversity

The purpose of this section is to introduce the notions of opinion, bias and diversity which we see as closely intertwined. We define an opinion as follows:

> **Opinion.** *An opinion is a subjective statement, i.e. a minimum semantically self-contained linguistic unit, asserted by at least one actor, called the opinion holder, at some point in time, but which cannot be verified according to an established standard of evaluation. It may express a view, attitude, or appraisal on an entity. This view is subjective, with positive/neutral/negative polarity (i.e. support for, or opposition to, the statement).*

By ‘entity’ we mean something that has a distinct, separate existence, not necessarily a material existence; it may be a concrete object or an abstract concept., In the sentence “President Obama said that police in Cambridge, Massachusetts, ‘acted stupidly’ in arresting a prominent black Harvard professor”, the opinion holder is *President Obama,* the statement is *police acted stupidly* which expresses an opinion of *negative* polarity. We then define bias as follows:

> **Bias.** *Bias is the degree of correlation between (a) the polarity of an opinion and (b) the context of the opinion holder.*

We thus see bias as a linking device. The polarity of an opinion is the degree to which a statement is positive, negative or neutral. The context may refer to a variety of factors, such as ideological, political, or educational background, ethnicity, race, profession, age, location, or time. Bias is potentially measurable directly in terms of a scale for this correlation e.g. measuring the minority/majority status of opinions in different contexts, particularly in relation to cultural diversity [2, 29, 30]. For example, by asking the question *What proportions of conservatives, liberals and socialists favoured integration of Turkey into the EU in 1999, 2004 and 2009?* we begin to see, in a scalable way, whether the polarity of an opinion is correlated with the particular context of the opinion holders and, indeed, whether changes in bias occur over time.

We define diversity in relation to definitions that have emerged from Media Content Analysis [27, 28, 29] as follows:

Diversity. *Diversity is the co-existence of contradictory opinions and/or statements (some typically non-factual or referring to opposing beliefs/opinions).*

There are various forms and aspects of diversity:

- The existence of opinions with different polarity about the same entity (subject), e.g., at different times;
- Diversity of themes, speakers, arguments, opinions, claims and ideas or frames;
- Diversity of norms, values, behaviour patterns, and mentalities;
- Diversity in terms of geographical (local, regional, national, international, global focus of information), social (between individuals, between and within groups), and systemic (organizational and societal) aspects in media content;
- Static (at one point in time) and dynamic (long-term) diversity;
- Internal diversity (within one source) and external diversity (between sources).

Generally speaking, the following **dimensions of diversity** can be distinguished, both in texts and images:

- Diversity of sources (multiplicity of sources of texts an images);
- Diversity of resources (e.g., images, text);
- Diversity of topic;
- Diversity of speakers/actors/opinion holders (e.g., variety of political affiliation of opinion holders)
- Diversity of opinions;
- Diversity of genre (e.g., blogs, news, comments);
- Diversity of language;
- Geographical/spatial diversity;
- Temporal diversity.

More specifically, dimensions of diversity in text can, at the very least, be distinguished at *document level* and at *statement level*. Specific dimensions of diversity can also be recognized for images. Besides the diversity of web content, *diversity in queries* may well be relevant, where possible dimensions include user intent, user gender, and the time of the formulation of the query.

3 State of the art

Key notions in our definition of bias are the polarity of an opinion and the context of the opinion holder. The definition of diversity depends crucially on opinion detection. Several techniques to identify and analyze opinion are available: **opinion mining,** or sentiment analysis, has been mainly considered as a binary or three-class classification problem. Applied techniques include natural language processing and machine learning [14] which are mostly applied to online product reviews. Some research explores the problem of identifying the opinion holder; Youngho et al. [12] exploit lexical and syntactic information; Kim and Hovy [13] analyze the semantic structure of sentences and use semantic-role labelling to label opinion holder and topic; Berthard et al. [11] propose a semantic parser-based system which identifies opinion propositions and opinion holders. In the latter system, the semantic parser labels sentences with thematic roles (e.g. Agent and Theme) by training statistical classifiers and is endowed with additional lexical and syntactic features to identify propositions and opinion holders. Work on relating opinion holders with their personal background is still unavailable. However, some techniques *do* consider diversity.

Diversity of search results in text retrieval has been considered in the context of result diversification. Since user queries may well be ambiguous as regards their intent, diversification attempts to find the right balance between having more relevant results of the 'correct' intent and having more diverse results in the top positions. In order to improve user satisfaction, the top N results are either ranked by diversity [15, 16] or diversified optionally by clustering them according to the different diversity dimensions covered [17, 18].

Diversity in queries is mostly related to user intent when posting a search query. Existing research in this area deals with classification of user queries according to content destination (e.g. informational, navigational, transactional) [31, 32]. Some values for diversity dimensions, as considered here, are certainly available through meta-information (e.g. source and resource dimensions). The identification of other values requires automatic algorithms for topic detection, language identification, information extraction and opinion mining.

Image search engines and the diversification of search results is a relatively new area of research, where one way of increasing diversity is to ensure that duplicate, or near-duplicate images in the retrieved set are hidden from the user [19], e.g. by forming clusters of similar images and showing one representative for each of them. Other approaches use semantic web technologies to help increase the diversity of the search results. For instance, in ImageCLEF [20] image search results are presented as columns corresponding to the individual topics discovered.

Context analysis identifies relevant information behind the content, especially spatial and temporal information. With images, such techniques can identify the original source of the picture which may be of better processing quality, or even for automatic tagging [21] (e.g. tags propagated from one image to another).

Diversity dimensions in images directly extracts values from EXIF information inserted automatically (digital camera) or manually (e.g. by the photographer) in an image file (jpeg format). In the absence of EXIF tags, some features can be derived using image retrieval techniques [22], forensic techniques [24, 25], and algorithms for automatically annotating images and extracting high-level semantic features [23].

4 The proposed framework

Our objective is to enhance the state of the art by developing search facilities that determine diversity in a completely automatic way and capture diversity in all its dimensions, whence the interdisciplinary approach. Below, we first introduce the methodologies we have identified, and then we show how they are combined in the framework (see Fig. 1) we propose:

- **Media Content Analysis (MCA)** from a social sciences perspective. The analysis typically starts from the formulation of some specific ***research questions***, in terms of topics, actors and patterns of interpretation (i.e. indications about how the discourses are framed) that need to be investigated. The work proceeds with the identification of specific ***variables*** (i.e. metadata), which make up the ***Codebook***. It consists of different characteristics for every variable to ask specifically about in the relevant media, and of the instructions for the manual coding. The set of relevant media (e.g. documents, newspapers, websites, blogs and forums) is called the ***document corpus*** (equal to sample in social sciences). In particular, variables are extracted on different levels of the documents: some address the whole document and its source, some focus on claims. Note that the term "claim" is taken from the recently-used method for analyzing public discourse (i.e. political claim analysis) and hence denotes "the expression of a political opinion by physical or verbal action in the public sphere" [1]. We refer to "claim" in a more general sense of "statement" as the expression of an opinion in the public sphere. The variables from the Codebook, which are further aggregated into ***indicators***, are used for statistical purposes when responding to research questions. The significance of this methodology lies precisely in its capacity to detect context and cultural diversity.
- **Multimodal Genre Analysis (MGA)** from a semiotic perspective. This two-step process first annotates parts of websites as text-and-image combinations i.e. multimodal ***meaning-making units***, and, then, as a set of ***hierarchical patterns*** (MGA templates) including, *inter alia*, genres and mini-genres such as logos, contact information, menus, 'running text' paragraphs. Detailed analysis of such patterns, functioning on different scalar levels, helps predict where specific information will or will not be found in a website. Inspired by Halliday's theory of meaning, which posits the existence of at least three separate meanings intertwined in every communicative act, this approach views opinion, bias, and other appraisal systems, as part of interpersonal meaning [2, 3] and not, in themselves, as part of what Halliday calls ideational meaning, i.e. the expression of ideas. In this view, language and other semiotic resources such as colour, gesture, gaze, shapes, lines etc. are pattern-forming systems which govern the relationship between interpersonal and ideational meaning-making systems. This approach thus has the potential to detect patterns and to predict where to find relevant information and opinions and bias vis-à-vis that information.
- **Facet Analysis (FA)** from a knowledge representation and organization perspective. FA is the process necessary for the construction of a ***Faceted Classification (FC)*** of a domain [4, 5]. An FC is basically a set of taxonomies, called ***facets***, which encode the knowledge structure of the corresponding domain in terms of the standard terms used, concepts and the semantic relations between them. Each facet may be said to encode a dimension of knowledge in that domain. For each domain, facets are grouped into specific fundamental categories. Originally, Ranganathan [4, 5] defined five fundamental categories: Personality, Matter, Energy, Space and Time (synthetically PMEST). Later on, Bhattacharyya [26] proposed a refinement which consists of four main categories, called DEPA: Discipline (D) (what we call a domain), Entity (E), Property (P)

and Action (A), plus another special category, called Modifier. For instance, in the medicine domain (D) the body parts (E), the diseases which affect them (P) and the actions taken to cure or prevent them (A) are clearly distinguished. Modifiers are used to sharpen the intention of a concept, e.g. “infectious disease”. An FC is typically used to classify books in the domain according to their specific meaning, in contrast with classical enumerative approaches. They have a well-defined structure, based on principles, and tend to encode shared perceptions of a domain among users, thus providing more organized input to semantics-based applications, such as semantic searching and navigation [8].

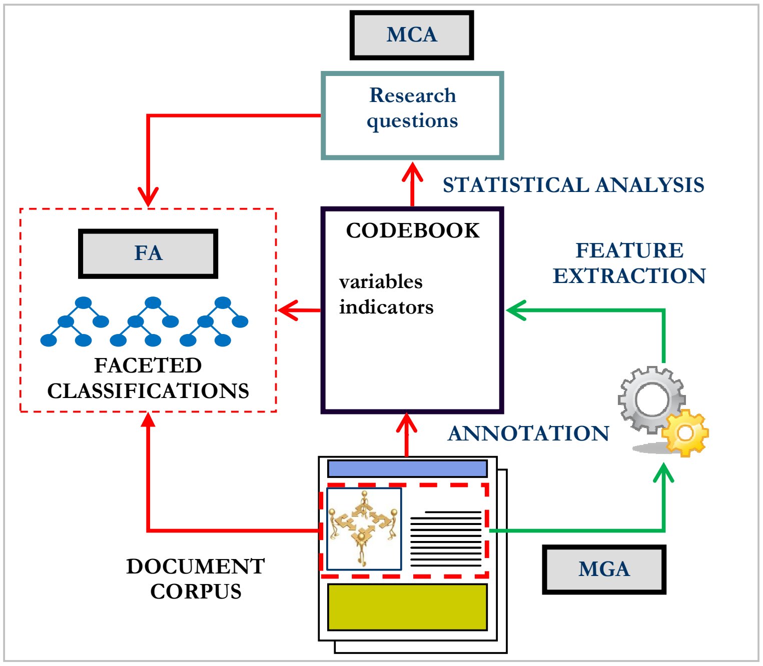

Fig. 1 – Technological integration of the methodologies contributing to the solution

Fig. 1 shows how these methodologies are integrated into an overall framework. Black boxes correspond to the methodologies described above. MCA is central. Most of the work in the project aims to automate this part of the process. Based on the analysis of typical research questions, Codebook templates and the document corpora used in MCA, FA is then used to generate FCs for the domains of interest. In order to test the framework, we chose the topic “European elections: migration, xenophobia, integration” as our use case and identified the following relevant domains: Political Science, Sociology, Psychology, Economics, Law, Geography, History, Philosophy, Religion and Information, Mass Media Research and Communication. Corresponding FCs have been generated. MGA contributes by identifying areas in the documents which are relevant to the extraction of specific information, for instance for opinions and bias. A set of feature extraction tools, which are meant to automate the annotation processes of the methodologies, are used to fill the Codebook with the information extracted from the document corpus. As described in [8], which extends the work in [6, 7], FCs are used as a controlled vocabulary during the whole process. For instance, the framework might parse content from different media and identify the main {concepts, people, political parties, countries, dates, resolutions, etc.} related to Xenophobia and which of them are the most {controversial, accepted, subjective, biased, etc.}.

5 Conclusions

In this paper we have described the key notions of opinion, bias and diversity, and the methodologies that, in our highly interdisciplinary approach, will synergically contribute to the development of advanced techniques for diversity-aware searching and navigation. The next challenges will mainly concern opinion, bias and diversity representation and management, automation of the annotation process and the implementation of the overall architecture.

Acknowledgments: In this paper we have presented and discussed some of the ideas and key notions which are the focus of the current LivingKnowledge EU FET project (http://livingknowledge-project.eu/), contract n. 231126 started the 1st February 2009. Thanks to all the other partners of the project for their support and precious feedback.

References

1. R. Koopmans, 2002. Codebook for the analysis of political mobilization and communication in European public spheres. Codebook from the Project: The Transformation of Political Mobilization and Communication in European Public Spheres. 5th Framework Program of the European Commission. Europub.com
2. M. Halliday, 1978. Language as Social Semiotic: The social interpretation of language and meaning. London: Edward Arnold.
3. M. Halliday, 1989. Part A. In M. Halliday, R. Hasan (eds.), Language, Context and Text: Aspects of language in a social-semiotic perspective. Oxford University Press, pp. 1-49.
4. S. R. Ranganathan, 1967. Prolegomena to library science. New York: Asia Publishing.
5. S. R. Ranganathan, 1962. Elements of library classification. New York: Asia Publishing.
6. F. Giunchiglia, M. Marchese, I. Zaihrayeu, 2006. Encoding Classifications into Lightweight Ontologies. Journal of Data Semantics 8, pp. 57-81, 2006.
7. F. Giunchiglia, I. Zaihrayeu, 2008. Lightweight ontologies. In S. LNCS, editor, Encyclopedia of Database Systems, 2008.
8. F. Giunchiglia, P. Shvaiko, M. Yatskevich, 2006. Discovering missing background knowledge in ontology matching. In ECAI, pp. 382–386, 2006.
9. F. Giunchiglia, B. Dutta, V. Maltese, 2009. Faceted Lightweight Ontologies. In “Conceptual Modeling: Foundations and Applications”, Alex Borgida, Vinay Chaudhri, Paolo Giorgini, Eric Yu (Eds.) LNCS 5600 Springer.
10. F. Giunchiglia, 2006. Managing Diversity in Knowledge. M. Ali and R. Dapoigny (Eds.): IEA/AIE 2006, LNAI 4031, p. 1, 2006. Springer-Verlag Berlin Heidelberg. Slides at: http://www.disi.unitn.it/~fausto/knowdive.ppt
11. S. Bethard, H. Yu, A. Thornton, V. Hatzivassiloglou, D. Jurafsky. Extracting opinion propositions and opinion holders using syntactic and lexical cues. Computing Attitude and Affect in Text: Theory and Applications, pp. 125-141, 2006.
12. Youngho Kim, Yuchul Jung, and Sung-Hyon Myaeng, 2007. Identifying opinion holders in opinion text from online newspapers. In GRC '07: Proceedings of the 2007 IEEE International Conference on Granular Computing, p. 699, Washington, DC, USA, 2007.

13. Soo-Min Kim, Eduard Hovy, 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. In Proceedings of the ACL Workshop on Sentiment and Subjectivity in Text, pp. 1-8. Association for Computational Linguistics, 2006.
14. Bo Pang, Lillian Lee, 2008. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2):1-135, 2008.
15. R. Agrawal, S. Gollapudi, A. Halverson, S. Ieong, 2009. Diversifying search results. In WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining, pp. 5-14, New York, NY, USA, 2009. ACM.
16. S. Gollapudi, A. Sharma, 2009. An axiomatic approach for result diversification. In WWW '09: Proceedings of the 18th international conference on World Wide Web, pp. 381-390, New York, NY, USA, 2009. ACM.
17. Fairspin, http://fairspin.org (last access: 2009/07/08)
18. Newssift, http://www.vewssift.com (last access: 2009/07/08)
19. T. Arni, P. Clough, M. Sanderson, M. Grubinger. Overview of the imageclefphoto 2008 photographic retrieval task. In Retrieved 18-06-2009, http://www.clef-campaign.org/2008/working notes/ImageCLEFphoto2008-¯nal.pdf, 200
20. ImageCLEF, http://www.imageclef.org (last access: 2009/07/08)
21. Tuffield, M., Harris, S., Dupplaw, D. P., Chakravarthy, A., Brewster, C., Gibbins, N., O'Hara, K., Ciravegna, F., Sleeman, D., Wilks, Y. and Shadbolt, N. R., 2006. Image annotation with Photocopain. In: 1st International Workshop on Semantic Web Annotations for Multimedia (SWAMM 2006) at WWW2006, pp. 22-26, 2006, Edinburgh, UK.
22. R. Datta, D. Joshi, J. Li, and J. Z. Wang, 2008. Image retrieval: Ideas, influences, and trends of the new age. ACM Comput. Surv. 40(2), pp. 1-60, 2008.
23. J. S. Hare and P. H. Lewis, 2005. On image retrieval using salient regions with vector-spaces and latent semantics. In CIVR, W. K. Leow, M. S. Lew, T.-S. Chua, W.-Y. Ma, L. Chaisorn, and E. M. Bakker, eds., LNCS 3568, pp. 540-549, Springer, 2005.
24. A. C. Popescu and H. Farid, 2004. Statistical tools for digital forensic. In Proc. 6th Int. Work. on Information Hiding, IH’04, Toronto, Canada, 2004, vol. 3200, pp. 128–147.
25. C. McKay, A. Swaminathan, H. Gou, and M. Wu, 2008. Image Acquisition Forensics: Forensic Analysis to Identify Imaging Source. In Proc. of IEEE Int. Conf. on Acoustic, Speech, and Signal Processing, ICASSP2008, Las Vegas, NV, 2008, pp. 1657-1660.
26. G. Bhattacharyya. POPSI: its fundamentals and procedure based on a general theory of subject indexing languages. Library Science with a Slant to Documentation, Vol. 16 No. 1, March, pp. 1-34, 1979.
27. D. McDonald, J. G./Dimmick, 2003. The Conceptualization and Measurement of Diversity. In Communication Research Vol. 30, No. 1/2003. S. pp. 60-79
28. D. McQuail, 2000. Mass communication theory. 4thed. London/Thousand Oaks/New Delhi: Sage
29. J. Van Cuilenburg, 2000. On Measuring Media Competition and Media Diversity. Concepts, Theories and Methods. In: Picard, Robert G. (Ed.): Measuring Media Content, Quality, and Diversity. Approaches and Issues in Content Research. Turku: Turku School of Economics. pp. 51-84
30. Paul J. Thibault, 2004. Brain, mind, and the signifying body. An ecosocial semiotic theory Continuum: London and New York. Foreword by Michael Halliday, p. 50
31. Gunther Kress (ed.) 1998 Communication and culture. An Introduction. Third Edition. Kensington: University of New South Wales Press.
32. R. Baeza-Yates, L. Calder´on-Benavides, C. Gonz´alez-Caro, 2006. The Intention Behind Web Queries. In Proceedings of String Processing and Information Retrieval (SPIRE 2006). Glasgow, Scotland, pp. 98-109.
33. Lee, U., Liu, Z. and Cho, J. 2005. Automatic Identification of User Goals in Web Search. In Proceedings of The World Wide Web Conference. Chiba, Japan, pp. 391-401.