Towards a Knowledge Diversity Model

Rakebul Hasan, Katharina Siorpaes,
Reto Krummenacher
Semantic Technology Institute (STI)
University of Innsbruck
A-6020 Innsbruck, Austria
firstname.lastname@sti2.at

Fabian Flöck
Institute of Applied Informatics and Formal
Description Methods
Karlsruhe Institute of Technology
D-76131 Karlsruhe, Germany
fabian.floeck@kit.edu

ABSTRACT

The Web is an unprecedented enabler for publishing, using and exchanging information at global scale. Virtually any topic is covered by an amazing diversity of opinions, viewpoints, mind sets and backgrounds. The research project RENDER works on methods and techniques to leverage diversity as a crucial source of innovation and creativity, and designs novel algorithms that exploits diversity for ranking, aggregating and presenting Web content. Essential in this respect is a knowledge model that makes accessible — cognitively to human users as well as computationally to the machine — the diversity in content. In this paper, we present a glossary of relevant terms that serves as baseline to the specification of the Knowledge Diversity Model.

Categories and Subject Descriptors

A.1 [**General Literature**]: Introductory and Survey; I.2.4 [**Computing Methodologies**]: Artificial Intelligence—*Knowledge Representation Formalisms and Methods*

Keywords

Knowledge diversity, Glossary, Knowledge model

1. INTRODUCTION

The Web is a tremendous facilitator and catalyst for the publication, use and exchange of information, fostering a global network of news, stories and statements which represent an amazing diversity of opinions, viewpoints, mind sets and backgrounds. Its design principles and core technology have led to an unprecedented growth in mass collaboration; a trend that is also increasingly impacting business environments.

The RENDER project[1] aims at leveraging the diversity inherently unfolding through world wide scale publishing and collaboration by developing methods, techniques, software and data sets that make diversity accessible as an important source of innovation and creativity, and by designing novel algorithms that reflect diversity in the ways information is selected, ranked, aggregated, presented and used.

An important component for the capturing of diversity in online documents, is a comprehensive knowledge model for representing diversity that reflects the plurality of opinions and viewpoints on a particular topic. In a first step, the considered content such as articles, blog entries or news feeds are transformed into a semantic representation according to the knowledge model that is accessible both cognitively to human users as well as computationally to the machine. The semantic representation is then leveraged for improving the selection and ranking of content, and the presentation to users. In RENDER, selection and ranking will go beyond widely adopted approaches based on popularity or personalization, and take opinions and viewpoints into account when computing the relevance of results.

In this paper we present a glossary of terms relevant in the scope of knowledge diversity. Creating a shared understanding of terms and relationships between terms is an essential first step towards the specification of a conceptual model for knowledge diversity. In that sense, this paper provides the necessary baseline for the definition of a knowledge diversity ontology, which allows for formalizing, gathering, evaluating and processing diversity in various (written) online medias.

In a first section (Section 2) we provide three motivating scenarios for this work, which are derived from the project's showcases that are brought to RENDER by Google, Wikimedia, and Telefonica. Section 3 provides a glossary of terms such as diversity, opinion, sentiment, bias and many more. Section 4 presents a short overview of the related work. In Section 5 we take a quick look at next steps, at how the targeted knowledge model will be used and leveraged in the given scenarios and throughout the project, and conclude the paper.

2. MOTIVATING SCENARIOS

In the following we present three motivating business scenarios for the formalization of a knowledge diversity model.

2.1 Wikipedia

Despite efforts for a balanced coverage at Wikipedia, systemic biases influenced by the individual views of the more than 100'000 volunteer contributors have been introduced. The increasing complexity of the control processes for creating and editing articles that are put in place to overcome the problem of biases, negatively impacts the growth of Wikipedia. Edit conflict resolution, arbitration committees, banning policies, a complex hierarchy of contributors, editors and administrators is not sustainable. Effectively, recent statistics show that the number of new articles has been decreasing dramatically over the past years, while the number of edits is still growing steadily. Discovering missing con-

[1] render-project.eu

DiversiWeb 2011, March 28, Hyderabad, India..

tent from one language version of Wikipedia to another, or the detection of diverse viewpoints within a topic or article are urgently needed support to the editorial team for managing and encouraging large-scale participation and sustainable growth. Diversity-empowered services such as quality or reliability assessment of an article or a specific statement, conflict resolution, anomaly detection, and cross-lingual consistency checking are expected to considerably improve the way information is currently managed in Wikipedia.

2.2 Google News

The news aggregator service of Google (Google News) indexes several ten thousands of news Web sites which are summaries into more than forty regional issues in more than 15 languages. The considered news content is created by professional journalists and by Web users, and offers as such a rich diversity of information. Current ranking algorithms result in news summaries that are dominated by popular viewpoints or opinion holders such as large news agencies. Alternative opinions, or arguments from smaller publishers often disappear and do not reach the interested audience. Consequently, even though Google aims for wide and comprehensive news coverage, the presented view points are highly biased. Manual processing is costly and impractical, and techniques to automatically discover diverse opinions, viewpoints and discussions surrounding a topic are required to fully leverage the richness in news content. Diversity-aware ranking of news posts for covering the most diverse view points on a particular topic, and enriching these with data from other sources like blogs, tweets, and wiki pages is expected to considerably increase the interconnection between diversifying news and discussions on the Web.

2.3 Customer Relationship Management

Telefónica is one of the World's largest telecommunications companies by market share, operating in 25 countries with a global customer base exceeding 280 millions. The company maintains various different communication channels including call centers, Web sites and public forums and blogs to collect customer feedback about their products and services. This offers a massive amount of valuable user opinions coming from diverse sources, countries and socio-demographic groups that are currently only marginally exploited, as the technical support for automation is missing and manual processing is not feasible to the desired extent. Discovering and automatically evaluating customer reactions and discussions are expected to allow Telefónica to react more efficiently and effectively to trends, to make more precise forecasts, and to eventually improve future business decisions.

3. KNOWLEDGE DIVERSITY GLOSSARY

The first step towards our knowledge diversity model is to create a shared understanding of the relevant terms and relationships between them in the scope of knowledge diversity. In this section, we present a summary of definitions of possibly relevant terms to get a rough understanding of the key concepts in the scope of knowledge diversity. We do not attempt to define these concepts in this paper; instead we refer to the existing definitions of these concepts.

Agent is described in DOLCE+DnS Ultralite as an agentive object, either physical (e.g. a person), or social (e.g. a corporation, an institution, a community).[2] As an extension of this concept, an agent expressing an opinion of his own can be called an *opinion holder*.

Belief is given by Wikipedia as "the psychological state in which an individual holds a proposition or premise to be true".[3] WordNet defines belief as "any cognitive content held as true", or alternatively as "a vague idea in which some confidence is placed".[4]

Bias is defined by Wikipedia as "an inclination to present or hold a partial perspective at the expense of (possibly equally valid) alternatives".[5] The definition of bias by Giunchiglia *et al.* in [5] states that "bias is the degree of correlation between (a) the polarity of an opinion and (b) the context of the opinion holder". The context can be a variety of factors such as ideological, political, or educational background, ethnicity, race, profession, age, location, or time.

Data is definded by WordNet as "a collection of facts from which conclusions may be drawn".[6] Wikipedia states that "the term data refers to qualitative or quantitative attributes of a variable or set of variables". Furthermore, data is the lowest level of abstraction from which first information and then knowledge are derived.[7]

Diversity is described in the philosophical sense, according to [3], as "the relation that holds between two entities when and only when they are not identical". In the Cambridge Advanced Learner's Dictionary diversity is defined as: "when many different types of things or people are included in something".[8] In [5] diversity is given from a more knowledge diversity focused point of view as "the co-existence of contradictory opinions and/or statements (some typically nonfactual or referring to opposing beliefs/opinions)". In the same paper, different dimensions of diversity are described such as: diversity of resources, diversity of topic, diversity of viewpoint, diversity of genre, diversity of language, geographical/spatial diversity, and temporal diversity.

Emotion is defined by Liu as "subjective feelings and thoughts" [7]. As Liu discusses, people use language expressions to describe their mental state (or feelings). According to [8], there are a large number of language expressions to depict the six types of emotions; i.e., *love*, *joy*, *surprise*, *anger*, *sadness* and *fear*. Similarly, people use a large number of opinion expressions to convey opinions with positive or negative sentiment.

Entity is described by Wikipedia as "something that has a distinct, separate existence, although it need not be a material existence".[9] In entity-relationship modelling, an entity is defined as "a thing which is recognized as being capable of an independent existence and which can be uniquely identified".[10]

[2] ontologydesignpatterns.org/ont/dul/DUL.owl

[3] en.wikipedia.org/wiki/Belief

[4] wordnetweb.princeton.edu/perl/webwn?s=belief

[5] en.wikipedia.org/wiki/Bias

[6] wordnetweb.princeton.edu/perl/webwn?s=data

[7] en.wikipedia.org/wiki/Data

[8] dictionary.cambridge.org/dictionary/british/diversity

[9] en.wikipedia.org/wiki/Entity

[10] en.wikipedia.org/wiki/Entity-relationship_model

Event is described in DOLCE+DnS Ultralite as "any physical, social, or mental process, event, or state". DOLCE+DnS Ultralite classifies events based on 'aspect' (e.g., stative, continuous, accomplishment, achievement, etc.), on 'agentivity' (e.g., intentional, natural, etc.), or on 'typical participants' (e.g., human, physical, abstract, food, etc.).

Fact, according to Liu, is the "objective expressions about entities, events and their properties" [7]. Wikipedia states that facts "refer to verified information about past or present circumstances or events which are presented as objective reality".[11] The Merriam-Webster Online Dictionary defines fact, *inter alia*, as 1) "the quality of being actual." 2) "something that has actual existence." or "An actual occurrence", 3. "a piece of information presented as having objective reality".[12]

Information is defined in [4] in terms of *data* + *meaning*:

σ is an instance of information, understood as semantic content, if and only if:

i) σ consists of n *data*, for $n \geqslant 1$;
ii) the data are *well formed*;
iii) the well-formed data are meaningful.

According to this definition, information is made of data and 'well formed' here means that data are rightly put together. Well formed and meaningful data are also known as *semantic content.* Information, understood as semantic content, has two major types: (a) *instructional* information, conveying the need for a specific action (b) *factual* information.

Information Object is described by DOLCE+DnS Ultralite as "a piece of information, such as a musical composition, a text, a word, a picture, independently from how it is concretely realized".

Knowledge is informally described in [2]. In a sentence like "John knows that Sara will come to the party", knowledge is "a relation between a knower, like John, and a proposition, that is, the idea expressed by a simple declarative sentence", like "Sara will come to the party". The proposition here are the abstract entities that can be *true* or *false*, right or wrong. More specifically, the sentences expressing the propositions, which are factual or non-factual, are *true* or *false.* The relationship between agents and propositions have different *propositional attitude* denoted by verbs like "know", "hope", "fear", "regret", and "doubt" etc. Brachman and Levesque do not consider the sentences involving knowledge that do not explicitly mention a proposition. For example, it is not clear if there is any useful proposition involved in the sentences like "John knows how to play guitar" or "John knows Bob well". Brachman and Levesque also discuss that the notion of *belief* is related to the notion of *knowledge.* People use the notion of *belief* if they do not want to claim that the judgement of an agent about the world is necessarily accurate.

Metadata is defined by Wikipedia as the "data providing information about one or more aspects of the data",[13]; e.g., means of creation of the data, purpose of the data, time and date of creation, creator or author of data, placement on a computer network where the data was created, or standards used. WordNet simplifies the meaning of metadata as "data about data".[14]

Object is described in DOLCE+DnS Ultralite as "any physical, social, or mental object, or a substance". The definition of objects by Liu states that "an object o is an entity which can be a product, person, event, organization, or topic [7]. It is associated with a pair, o: *(T, A)*, where T is a hierarchy of components (or parts), sub-components, and so on, and A is a set of attributes of o. Each component has its own set of sub-components and attributes".

Objectivity is the expression of facts [1]. Wikipedia moreover describes objectivity as "a proposition is generally considered to be objectively true when its truth conditions are mind-independent – that is, not the result of any judgements made by a conscious entity or subject".[15] WordNet defines it as the "judgment based on observable phenomena and uninfluenced by emotions or personal prejudices",[16] while according to [7] objective sentences express factual information about the world.

Object Feature represents the components and attributes of objects [7]. The term object feature is also referred simply as feature. Object features are used to simplify the complexity of hierarchical representation of the components of objects.

Opinion is defined by Wikipedia as "a subjective statement or thought about an issue or topic, and is the result of emotion or interpretation of facts".[17] Furthermore, "an opinion may be supported by an argument, although people may draw opposing opinions from the same set of facts". In [5], opinion is defined as "a 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)". Another definition of opinion, given by Liu [7], states that "an opinion on a feature f is a positive or negative view, attitude, emotion or appraisal on f from an opinion holder".

Opinion Expression is given by Liu as subjective expression that describes sentiments, appraisals or feeling toward entities, events and their properties [7]. More generally speaking, it could be said that opinion expressions are individual statements that contain an assessment of reality from the point of view of the *opinion holder.*

Opinion Holder, according to Liu [7], is "the person or organization that expresses the opinion"; see *Agent* above.

Polarity of Opinion on a feature f indicates if the opinion is positive, negative or neutral [7]. [5] describes polarity as the degree to which a statement is positive, negative or neutral. The polarity of an opinion is also known as sentiment orientation or semantic orientation [7].

Sentiment is defined in the American Heritage Dictionary

[11] `en.wikipedia.org/wiki/Fact`

[12] `www.merriam-webster.com/dictionary/fact`

[13] `en.wikipedia.org/wiki/Metadata`

[14] `wordnetweb.princeton.edu/perl/webwn?s=metadata`

[15] `en.wikipedia.org/wiki/Objectivity_(philosophy)`

[16] `wordnetweb.princeton.edu/perl/webwn?s=objectivity`

[17] `en.wikipedia.org/wiki/Opinion`

of the English Language as "a thought, view, or attitude, especially one based mainly on emotion instead of reason".[18] Sentiments can be seen as a way to express opinions. Hence, sentiments, as much as opinions, can be negative, positive or neutral [7].

Subjectivity refers to the subject and the perspective, feelings, beliefs, and desires of the subject [6]. Liu defines subjective sentences as the sentences which "express some personal feelings or beliefs" [7].

Text is defined by Dictionary.com, in the linguistic sense, as "a unit of connected speech or writing, especially composed of more than one sentence, that forms a cohesive whole".[19] The Free On-line Dictionary of Computing describes it as the "textual material in the mainstream sense", and in the computing sense as the "data in ordinary ASCII or EBCDIC representation", where ASCII and EBCDIC are computer codes for representing alphanumeric characters.[20]

Topic has three definitions in Wikipedia: "a.) the phrase in a clause that the rest of the clause is understood to be about, b.) the phrase in a discourse that the rest of the discourse is understood to be about, c.) a special position in a clause (often at the right or left-edge of the clause) where topics typically appear".[21] WordNet defines topic as "the subject matter of a conversation or discussion".[22]

4. RELATED WORK

Giunchiglia *et al.* consider knowledge diversity as an asset to improve navigation and search [5], however, they do not provide a representation model to represent the knowledge gathered using their technology. Liu introduces the core topics in the field of sentiment analysis and opinion mining, such as sentiment and subjectivity classification, feature-based sentiment analysis, sentiment analysis of comparative sentences, opinion search and retrieval, opinion spam and utility of opinions [7]. Liu provides definitions of the relevant concepts but the work is aimed at the processing of opinions, and not at representing opinions. Balahur and Steinberger provide their insight on sentiment analysis for the news domain [1], and as such argue the need for clearly defining the source and target of a sentiment. They provide guidelines on annotating news contents with different sentiments, however, they do neither discuss the representation of the captured knowledge.

The listed works present technologies and methodologies to gather different aspects of diversity, but they do not provide any representation model for this gathered knowledge. In contrast, our aim is to work towards developing a knowledge diversity model to represent the different aspects of diversity.

5. FUTURE WORK AND CONCLUSIONS

The goal of this paper was to collect a comprehensive glossary of terms that are relevant in the context of knowledge diversity. Aspects such as opinion, sentiment or bias are essential in understanding the diversity of news posts, Wikipedia articles, or customer feedback. Only when diversity can be computationally accessible to the machine, the capturing and interpretation of opinions and sentiments can be automated and results extracted at larger scale.

The intention is to derive a knowledge diversity model from the glossary presented in this paper. In the next step it will be necessary to determine the concrete questions that will have to be answered for the showcase scenarios, and to extract the definitions that cover these relevant aspects. Another important future work would be to determine the relationships among the aforementioned concepts. As an example, based on the definition presented in this paper we can conclude that sentiments are a way to express opinions. Subjectivity refers to the perspective, beliefs and feelings of a person. Bias is influenced by someone's personal opinion. A particular bias can influence the subjectivity of a sentence when it contains an opinion. Opinions are expressed by the opinion expressions. Opinion expressions are subjective statements contained in the information objects. The concepts and relationships can be seen as the baseline for the specification of the knowledge diversity ontology that yields the schema information for semantically capturing the diversity and context of the textual content considered. Context, also not part of the collected definitions above, is important to interpret diverse standpoints in view of their socio-demographic, spatio-temporal and historic relationship to each other. In many situation, taking the customer relationship management as an example, it is not only relevant to interpret diverging opinions and sentiments of customers but also to understand the situation of the opinion holders such as for example their country of residence. This allows for drawing further conclusions relevant for shaping the business.

Acknowledgments: The work presented in this paper is supported by the European Union's 7th Framework Programme (FP7/2007-2013) under Grant Agreement 257790.

6. REFERENCES

[1] A. Balahur and R. Steinberger. Rethinking Sentiment Analysis in the News: from Theory to Practice and back. In *1st Workshop on Opinion Mining and Sentiment Analysis*, 2009.

[2] R. Brachman and H. Levesque. *Knowledge representation and reasoning.* Morgan Kaufmann Publishers, 2004.

[3] J. Butterfield. *Collins English dictionary: Complete and unabridged.* HarperCollins Publishers, 2003.

[4] L. Floridi. *Information: A Very Short Introduction.* Oxford University Press, 2010.

[5] F. Giunchiglia, V. Maltese, D. Madalli, A. Baldry, C. Wallner, P. Lewis, K. Denecke, D. Skoutas, and I. Marenzi. Foundations for the representation of diversity, evolution, opinion and bias. Technical Report DISI-09-063, University of Trento, 2009.

[6] T. Honderich. *The Oxford Companion to Philosophy.* Oxford University Press, 2005.

[7] B. Liu. *Handbook of Natural Language Processing*, chapter Sentiment Analysis and Subjectivity, pages 627–666. CRC Press, 2010.

[8] W. Parrott. *Emotions in Social Psychology: Essential Readings.* Psychology Press, 2001.

[18] www.houghtonmifflinbooks.com/ahd/

[19] dictionary.reference.com/browse/text

[20] foldoc.org/text

[21] en.wikipedia.org/wiki/Topic

[22] wordnetweb.princeton.edu/perl/webwn?s=topic