Knowledge-Based Systems xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Enriching semantic knowledge bases for opinion mining in big data applications

A. Weichselbraun [a], S. Gindl [b,*], A. Scharl [b]

[a] University of Applied Sciences Chur, Switzerland
[b] MODUL University Vienna, Austria

ARTICLE INFO

Article history:
Received 22 November 2013
Received in revised form 19 April 2014
Accepted 26 April 2014
Available online xxxx

Keywords:
Opinion mining
Contextualization
Commonsense

ABSTRACT

This paper presents a novel method for contextualizing and enriching large semantic knowledge bases for opinion mining with a focus on Web intelligence platforms and other high-throughput big data applications. The method is not only applicable to traditional sentiment lexicons, but also to more comprehensive, multi-dimensional affective resources such as SenticNet. It comprises the following steps: (i) identify ambiguous sentiment terms, (ii) provide context information extracted from a domain-specific training corpus, and (iii) ground this contextual information to structured background knowledge sources such as ConceptNet and WordNet. A quantitative evaluation shows a significant improvement when using an enriched version of SenticNet for polarity classification. Crowdsourced gold standard data in conjunction with a qualitative evaluation sheds light on the strengths and weaknesses of the concept grounding, and on the quality of the enrichment process.

1. Introduction

Communication experts and decision makers aim to understand how stakeholders perceive their announcements and actions, and how news coverage and social media channels affect these perceptions. To address these questions, this article describes the integration and automated extension of semantic knowledge repositories. Building upon a novel approach to contextualized sentiment analysis [19], we introduce methods that can ground and enrich identified concepts. This integration of semantic knowledge repositories is an important stepping stone towards making sense of big data. Extracting factual and affective knowledge from these repositories will provide a deeper understanding of opinions expressed in user-generated content from social media platforms, news articles, scientific publications, etc.

The knowledge extraction tools that analyze the Social Web typically provide frequency and sentiment metrics on either a document or sentence level. Sentiment is an important and insightful indicator. However, even when it is measured accurately, this single metric often cannot address fundamental questions posed by decision makers. Communication experts who are responsible for marketing and public outreach campaigns, for example, want to know if their message reaches intended groups, how their communication strategy impacts observable patterns in online coverage, and which portion of the identified sentiment actually refers to their organization. The U.S. National Oceanic and Atmospheric Administration (NOAA) is a good example. The NOAA Climate Program Office has adopted the authors' previous work on opinion mining as an essential part of its online evaluation strategy. Fig. 1 shows a screenshot of the system, which is based on the webLyzard big data and Web intelligence platform [15]. The dashboard uses color coding to embed sentiment information into various interface components including a relevance-ranked list of search results, trend charts, and a range of other interactive visualizations such as tag clouds, keyword graphs, word trees and geographic maps. Communication experts at NOAA use the system to track whether social media users associate NOAA with "climate change", for example, which is an important aspect of their communication and outreach goals. With regard to sentiment analysis, this poses an interesting challenge because the term "climate change" typically carries a negative connotation. In such cases, it is imperative to differentiate the sentiment of concepts that are merely co-referenced in a document ("NOAA", "climate change"), and opinions that are directed towards an organization.

User-generated product reviews are another example illustrating the importance of identifying specific opinion targets when analyzing the Social Web. Users tend to comment not only on a product in general (e.g., digital camera), but also on its various

* Corresponding author. Tel.: +43 1 3203555 525.
E-mail addresses: albert.weichselbraun@htwchur.ch (A. Weichselbraun), stefan.gindl@modul.ac.at (S. Gindl), arno.scharl@modul.ac.at (A. Scharl).

http://dx.doi.org/10.1016/j.knosys.2014.04.039
0950-7051/© 2014 Published by Elsevier B.V.

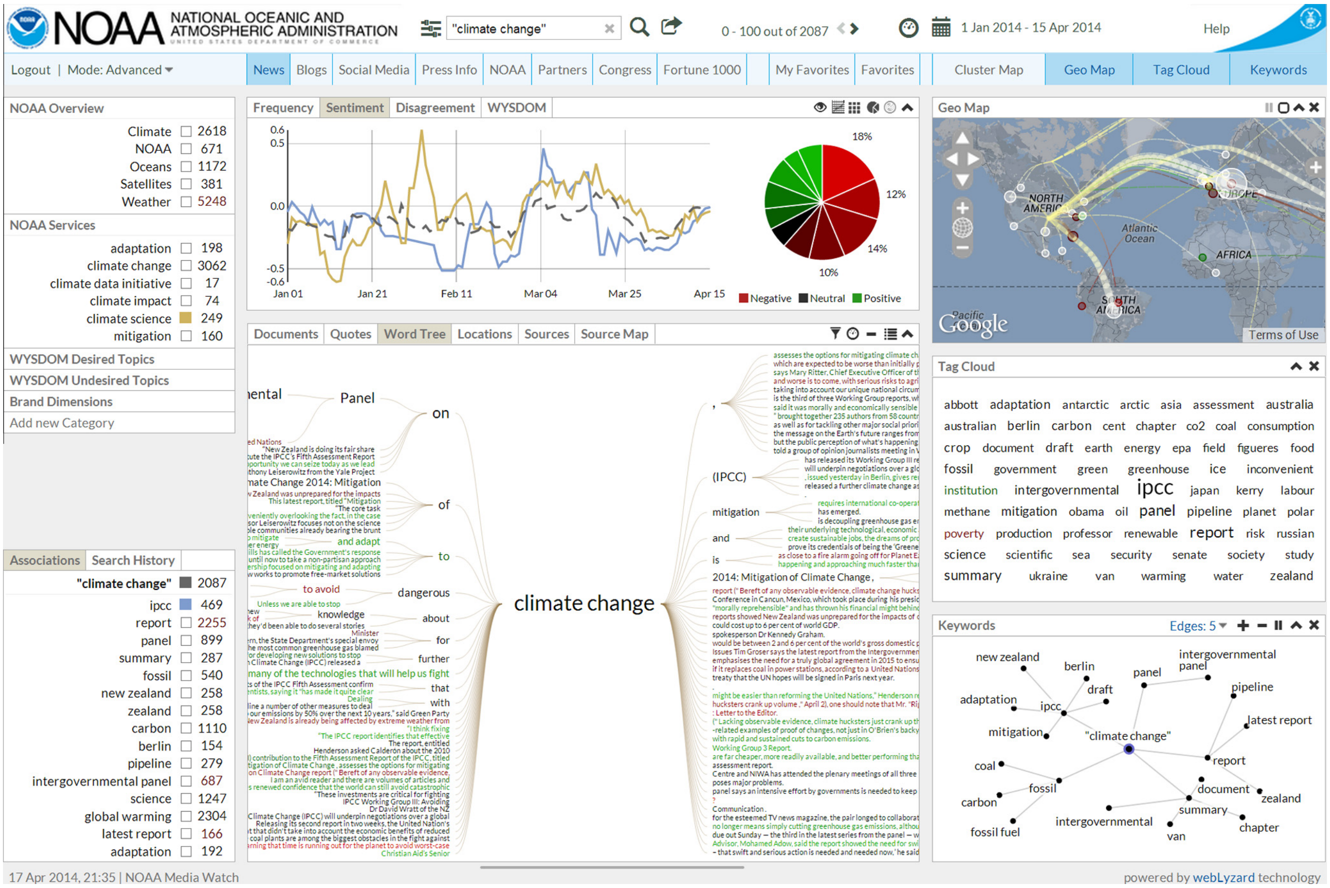

Fig. 1. Screenshot of a Web intelligence portal built for the NOAA Climate Program Office, showing results for a query on "climate change" based on news media coverage between January and April 2014.

aspects – shutter speed, quality of the lens, retail price, etc. The observable preference of users to compare features across products rather than assessing them in isolation underscores the need for flexible approaches to concept grounding and enrichment, which are granular enough to distinguish between the specific aspects of an entity. For evaluation purposes, therefore, this paper uses reviews from Amazon.com about electronics and software products as well as reviews from the Internet Movie Database (www.imdb.com) in the categories comedy, crime and drama.

2. Related work

Many opinion mining tools rely on sentiment lexicons as linguistic resources that attach polarity values and strengths to sentiment terms. Static polarity values may serve as a good baseline, but a closer examination reveals the need for more differentiated approaches. Cambria and White emphasize the need for a shift from simple syntactic (bag-of-words) approaches to semantic (bag-of-concept) or even pragmatic (bag-of-narratives) ones in their extensive review on natural language processing [5]. Depending on the context, a term might lose its opinionated characteristic, or its polarity might change – e.g., "good" expressing a positive emotion versus "good" as the cargo of a freight train.

Gangemi et al. [9] emphasize the importance of sentiment contextualization as one of seven major challenges in the area of opinion holder and target detection. Existing approaches handle it in different ways – e.g., by vector space modeling [6], by invoking language models [11], or by applying sentence- and discourse-based context shifters [20], rule-based approaches [7] or linguistic patterns [21]. Xia et al. [22] address the problem of contextual polarity change by employing an ensemble of part-of-speech (POS) features combined with a sample selector. The sample selector uses principal component analysis to select samples from the source domain that are similar to the target domain. Enriching sentiment lexicons with context knowledge is another research avenue being pursued [12]. Gindl et al. [10] separate ambiguous sentiment terms from terms with stable polarity, a process that yields contextualized sentiment lexicons. Embedding context information into the lexicon allows adapting an ambiguous term's polarity if the context indicates a polarity shift.

Structured knowledge contained in external linguistic repositories can support this contextualization process. Efforts to extend the well-known WordNet repository [8] have resulted in language resources such as SentiWordNet [1] and WordNetAffect [17]. The former attaches objectivity and polarity values to WordNet synsets, while the latter enriches WordNet with labels for affective categories. Tsai et al. [18] present another approach to enriching language resources. They apply iterative regression and a random walk strategy to label ConceptNet [16] elements with sentiment values. Poria et al. [13] merge SenticNet [3] and WordNetAffect to provide emotive labels for SenticNet. SenticNet itself uses ConceptNet by blending its knowledge with WordNetAffect, and inferring concept polarities from the Hourglass of Emotion [2].

Our previous work extended this line of research by accomplishing cross-domain contextualization [10]. Complementing related work that applies common and common-sense knowledge to improve sentiment analysis [4], this paper specifically targets the problem of correctly interpreting ambiguous sentiment terms. We ground such terms depending on their actual usage to

Q1 Please cite this article in press as: A. Weichselbraun et al., Enriching semantic knowledge bases for opinion mining in big data applications, Knowl. Based Syst. (2014), http://dx.doi.org/10.1016/j.knosys.2014.04.039