WEB-BASED KNOWLEDGE EXTRACTION AND THE COGNITIVE CHARACTERIZATION OF CULTURAL GROUPS

Antonio Penta*, Paul R. Smart*, Winston R. Sieck^, and Nigel R. Shadbolt*
*School of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, UK
{ap7, ps02v, nrs}@ecs.soton.ac.uk
^Applied Research Associates, Inc., 1750 Commerce Center Blvd, N., Fairborn, OH, 45324
wsieck@ara.com

ABSTRACT

The advent of Web 2.0 has provided new opportunities for cultural analysts to understand more about the cognitive characteristics of cultural groups. In particular, user-contributed content provides important indications as to the beliefs, attitudes and values of cultural groups, and this is an important focus of attention for those concerned with the development of cognitively-relevant models. In order to support the exploitation of the Web in the context of cultural modeling activities, it is important to deal with both the large-scale nature of the Web and the current dominance of natural language formats. In this paper, we outline an approach to support the exploitation of the Web in the context of cultural modeling activities. The approach begins with the development of qualitative cultural models (which describe the beliefs, concepts and values of cultural groups), and these models are subsequently used to develop an ontology-based information extraction capability (which harvests model-relevant information from online textual resources). We are currently developing a system to support the approach, and the continued development of this system should enable cultural analysts to more fully exploit the Web for the purpose of developing more accurate, detailed and predictively-relevant cognitive models.

PRIMARY TRACK

Social-Cultural Data

SECONDARY TRACK

Understanding and Modeling Human Behavior

DESCRIPTION

The Web is a potentially valuable source of culture-relevant information, and it is an obvious focus of attention for those interested in developing a better understanding of the characteristics of various cultural groups. In addition to providing information about the geographic location, history and demographics of cultural groups, the advent of Web 2.0 (which is characterized by greater levels of user participation in the creation, maintenance and editing of online content) has provided new opportunities for cultural analysts to understand more about the cognitive characteristics of cultural groups. Cognitive characteristics include the beliefs, attitudes and values that are shared by group members, and a wide range of user-contributed resources (including blogs, videos, discussion forums, news broadcasts, and organizational websites) provide information about these characteristics.

Understanding the cognitive characteristics of cultural groups is important because of the role such characteristics play in influencing the behavior of group members. Thus, if we want to understand (explain and predict) the behavior of group members, it often makes sense to focus attention on the beliefs, attitudes and values that are held by those members. One approach that explicitly embraces this idea is proposed by Sieck et al [1]. They suggest that the development of ‘cultural models’ (which represent the beliefs and values of cultural groups) can lead to a better understanding of the reasons why particular decision outcomes are sanctioned by group members. An example of such a model is presented in Figure 1. This model, which was developed from articles describing jihadist narratives, illustrates the beliefs held by extremist Sunni Muslims regarding current socio-political relationships between Islam and the West. The cultural model illustrates the concepts shared by the group, as well as their common knowledge of the causal relationships between those concepts. The nodes in the model represent simple concepts, such as “Muslim Honor”, with the shape assigned to each node indicating whether a positive (circle) or negative (hexagon) evaluation applies to the concept. The arrows represent causal beliefs, with the +/- symbols indicating the polarity of the causal relationship.

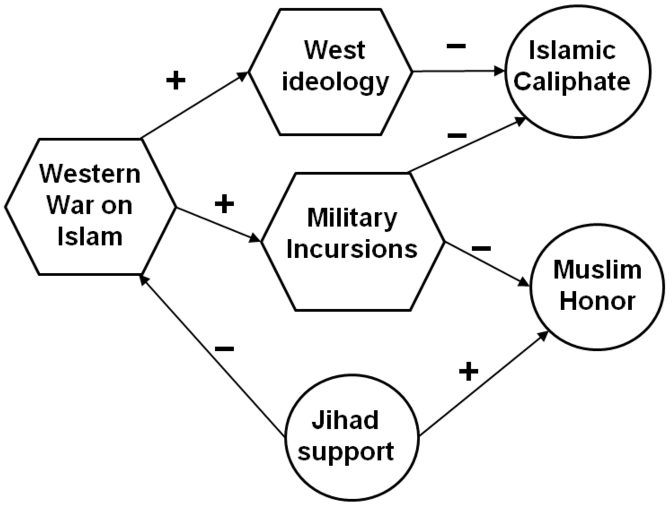

Figure 1. Simplifed cultural model of extremist Sunni Muslims beliefs about current socio-political relationships between Islam and the West.

The development of cultural models obviously requires access to resources indicating the beliefs and values of group members, and it is here that the potential value of the Web to support cultural model development becomes apparent. Inasmuch as we can leverage the latent potential of the Web to yield information about the cognitive characteristics of cultural groups, then cultural analysts should be better enabled to develop more accurate, detailed and predictively-relevant cultural models.

In order to enable cultural analysts to exploit the Web in the context of cultural modeling activities, we are developing a system to support the automated extraction of culture-relevant information from online textual resources [see 2]. This process of automated knowledge extraction is important for a number of reasons. Firstly, the large-scale nature of the Web and the current dominance of natural language formats makes culture-relevant information both difficult to find and difficult to process. By developing state-of-the-art knowledge extraction capabilities, we hope to support cultural analysts in identifying, classifying and processing relevant online resources. Secondly, a key part of cultural modeling is the transition from what are called ‘qualitative cultural models’ to ‘quantitative cultural models’ [see 1]. Qualitative cultural models merely seek to identify the concepts, beliefs and values associated with a particular cultural group; quantitative cultural models, in contrast, seek to supplement qualitative models with information about the prevalence of each concept, belief and value in the target population. The use of Web-based knowledge extraction techniques is a potentially important element of quantitative model development: it enables cultural analysts to estimate the relative frequency of model elements in a target population and thus develop quantitative extensions of qualitative cultural models.

We have identified a number of steps associated with the use of Web-based knowledge extraction technology to support the process of cultural model development. The first step in the process is to develop an initial qualitative cultural model using a limited set of knowledge sources [see 1]. The second step involves the development of a cultural ontology using the qualitative cultural model as a reference point. This ontology is represented using the Ontology Web Language (OWL), which has emerged as a de facto standard for formal knowledge representation on the World Wide Web. The third step is to manually annotate sample texts using the cultural ontology in order to provide a training corpus for rule learning. Rule learning, in the current context, is mediated by the $(LP)^2$ algorithm, which is a supervised algorithm that has been used to develop a variety of adaptive information extraction and semantic annotation capabilities [3, 4]. Following the development of information extraction rules, the rules are then applied to Web resources in the fourth step in order to identify concepts defined in the initial qualitative cultural model. Step five consists in the identification and extraction of causal relationships corresponding to beliefs in the cultural model. The extraction of causal relationships is a difficult challenge because information extraction techniques tend to focus on the extraction of particular entities in a text, rather than the relationships between the entities. We attempt to extract causal relationships using an approach that combines the use of background knowledge in the form of a domain ontology with the general purpose lexical database, WordNet [see 2 for more details]. Finally, in step 6, the extracted cultural knowledge is integrated, stored, and used to estimate the relative frequencies of the various elements of the initial qualitative cultural model.

We are currently developing a system to support this overall process in the context of the IEXTREME project, which is a trans-Atlantic project, funded by the U.S. Office of Naval Research [see 2]. By implementing Web-based knowledge extraction capabilities that are sensitive to the interests, concerns and requirements of cultural modelers, we aim to support cultural analysts in developing a better understanding of disparate cultural groups.

ACKNOWLEDGEMENTS

This research was supported by Contract N00014-10-C-0078 from the U.S. Office of Naval Research.

BIOGRAPHY

Dr Antonio Penta is a research fellow in the School of Electronics and Computer Science at the University of Southampton. He currently works on the IEXTREME project under the supervision of Professor Nigel Shadbolt and Dr Paul Smart. Previously, he was a post-doctoral researcher at the Politecnico of Turin and research fellow in the Department of Computer Science at the University of Naples. He received a PhD in Computer Science from the University of Naples in 2008 under the supervision of Professor Antonio Picariello. He has also been a visiting researcher at the Politecnico of Milan, Rensselaer Polytechnic Institute, and the

University of Maryland. His research interests include knowledge representation for multimedia data; multimedia information processing, extraction, integration and reasoning; multimedia databases; 3D databases and e-Government.

REFERENCES

[1] W. R. Sieck, L. Rasmussen, and P. R. Smart, "Cultural Network Analysis: A Cognitive Approach to Cultural Modeling," in Network Science for Military Coalition Operations: Information Extraction and Interaction, D. Verma, Ed. Hershey, Pennsylvania, USA.: IGI Global, 2010.

[2] P. R. Smart, "Development of an Web-Based Knowledge Extraction System to Support Cultural Modelling and Analysis," School of Electronics and Computer Science, University of Southampton, Southampton, England, Technical Report IEXTREME/WBKE_Report, 2010.

[3] M. Vargas-Vera, E. Motta, J. Domingue, M. Lanzoni, A. Stutt, and F. Ciravegna, "MnM: ontology driven semi-automatic or automatic support for semantic markup," in 13th International Conference on Knowledge Engineering and Knowledge Management, Siguenza, Spain, 2002.

[4] F. Ciravegna and Y. Wilks, "Designing adaptive information extraction for the Semantic Web in Amilcare," in Annotation for the Semantic Web, S. Handschuh and S. Staab, Eds. Amsterdam: IOS Press, 2003.