Using Web-Based Knowledge Extraction Techniques to Support Cultural Modeling

Paul R. Smart[1], Winston R. Sieck[2], and Nigel R. Shadbolt[1]

[1] School of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, UK,
{ps02v, nrs}@ecs.soton.ac.uk
[2] Applied Research Associates, Inc., 1750 Commerce Center Blvd, N., Fairborn, OH, 45324, USA,
wsieck@ara.com

Abstract. The World Wide Web is a potentially valuable source of information about the cognitive characteristics of cultural groups. However, attempts to use the Web in the context of cultural modeling activities are hampered by 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 for 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. Our approach represents an attempt to combine conventional approaches to information extraction with epidemiological perspectives of culture and network-based approaches to cultural analysis. The approach can be used, we suggest, to support the development of models providing a better understanding of the cognitive characteristics of particular cultural groups.

Keywords: cultural modeling, ontology-based information extraction, culture, cognition, knowledge extraction, world wide web

1 Introduction

The World Wide Web (WWW) is a valuable source of culture-relevant information, and it is therefore an important resource for those interested in developing cultural models. The exploitation of the WWW in the context of cultural modeling is, however, hampered both by the large-scale nature of the Web (which makes relevant information difficult to locate) and the current dominance of natural language formats (which complicates the use of automated approaches to information processing). In this paper, we describe an approach to support the use of the Web in cultural modeling activities. The approach is based on the development of ontology-based information extraction capabilities, and it combines the use of Semantic Web technologies and natural language processing (NLP) techniques with an epidemiological perspective of culture [1] and the use

of network-based approaches to cultural analysis [2]. Technological support for the approach is currently being developed in the context of the IEXTREME[1] project, which is funded by the U.S. Office of Naval Research.

The structure of the paper is as follows. In Section 2, we outline what is meant by the term 'culture', and we describe an approach to cultural modeling that is based on the development of models representing the ideas associated with particular cultural groups. In Section 3, we describe our approach to the development of Web-based knowledge extraction capabilities to support cultural model development. This approach combines conventional approaches to information extraction with semantically-enriched representations of cultural models, and it seeks to provide a culture-oriented ontology-based information extraction (OBIE) capability for the WWW.

2 Cultural Models and Cultural Network Analysis

Before addressing the use of the Web to study culture, we first need to define what is meant by the term 'culture'. As is to be expected in any highly interdisciplinary field, there are a variety of conceptions of culture. Our conception is distinctly cognitive in nature, and it is based on an epidemiological perspective [1]. A fundamental assumption of this perspective is that shared developmental experiences lead to important similarities in the mental representations (e.g. values and causal knowledge) that are distributed among members of a population. Culturally widespread ideas ground the distribution of behavioral norms, discussions, interpretations, and affective reactions, and researchers working within the epidemiological perspective thus seek to describe and explain the prevalence and spread of ideas within specific populations.

Working from this perspective, we previously developed a technique called cultural network analysis (CNA), which is a method for describing the 'ideas' that are shared by members of cultural groups [2]. CNA discriminates between three kinds of ideas, namely, concepts, values, and causal beliefs, which together constitute the contents of what are called 'cultural models'. These cultural models typically rely on the use of belief network diagrams to show how the set of relevant ideas relate to one another (see Fig. 1 for an example).

In general, we can distinguish two types of cultural models: qualitative and quantitative cultural models (see [2]). Qualitative cultural models present the ideas associated with a particular group, whereas quantitative models add information about the prevalence of those ideas in the target population. In addition to seeing the approach described in this paper as a means to validate and refine qualitative cultural models, it is also possible to see the approach as enabling a cultural analyst to estimate the relative frequency of ideas in a target population and thus develop quantitative extensions of qualitative cultural models (see Section 3.6).

[1] See `http://www.ecs.soton.ac.uk/research/projects/746`.

3 Web-Based Knowledge Extraction for Cultural Model Development

In this section, we describe an approach to cultural model development that combines CNA with state-of-the-art approaches to knowledge representation and Web-based information extraction. The aim is to better enable cultural model developers to exploit the WWW as a source of culture-relevant information. The approach we describe is based on a decade of research into OBIE systems (see [3] for a review), and it combines conventional approaches to information extraction with an ontology that provides background knowledge about the kinds of entities and relationships that are deemed important in a cultural modeling context. The first step in the process is to develop an initial qualitative cultural model using a limited set of knowledge sources (see [2] for more details on this step). 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 WWW. 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 [4, 5]. Following the development of information extraction rules, the rules are then applied to Web resources in the fourth step in order to identify instances of the entities defined in the initial qualitative cultural model. Step five consists in the identification and extraction of causal relations. The extraction of causal relationships is a difficult challenge because techniques for information extraction have tended to focus on the extraction of particular entities in a text, rather than the relationships between those 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 [6]. Finally, in step six, the extracted cultural knowledge is integrated, stored, and used to estimate the relative frequencies of the various ideas presented in the initial qualitative cultural model. We briefly describe each of these steps in subsequent sections.

3.1 Step 1: Develop Qualitative Cultural Model

The technique used to develop qualitative cultural models has been described in previous work [2], and we will not reiterate the details of the technique here. Fig. 1 illustrates a simplified qualitative cultural model that represents an extremist Sunni Muslim's beliefs about current socio-political relationships between Islam and the West. The set of ideas represented in Fig. 1 were extracted from articles that describe jihadist narratives, and they are presented here for illustrative purposes. The cultural model illustrates concepts shared by the group, as well as their common knowledge of the causal relationships between those concepts. This shared knowledge influences expectations about how socio-political

relationships will unfold, and it provides a basis for the selection of particular actions and decision outcomes; for example, the decision to support jihad.

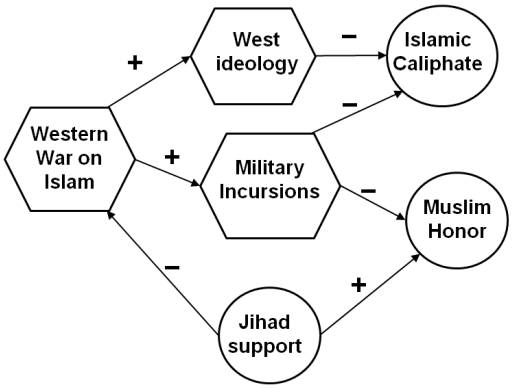

Fig. 1. Sunni extremist cultural model of jihad (simplified).

Fig. 1 shows the three different kinds of ideas that are the targets of a culture-oriented knowledge extraction system. These ideas include simple concepts such as "Western ideology" and "Muslim Honor", each represented as closed shapes in Fig. 1. They also include causal beliefs; for example, the idea that Western ideology (e.g. secularism, nationalism) is inhibiting the formation of a unified Islamic caliphate and the idea that the West promotes this ideology because it is engaged in a covert war against Islam. These causal beliefs are represented as arrows in the figure, with the +/- symbols indicating the polarity of the causal relationship. Finally, Fig. 1 portrays values using specific shapes, with circles indicating "positive" outcomes and hexagons indicating "negative" outcomes. Developing an Islamic caliphate is thus a good thing according to the cultural model. Maintaining (and enhancing) Muslim honor is likewise valued. According to the model, jihad is viewed positively and should be supported by the model's adherents due to the perceived anticipated consequences for Muslims. Most directly, support for jihad decreases the chances that the West will continue its war against Islam, and it enhances collective Muslim honor.

3.2 Step 2: Develop Cultural Ontology

Once an initial qualitative cultural model has been developed, the next step in the process is to develop an ontology that represents the contents of the model. The main reason this step is undertaken is that it enables the cultural model to be used to support information extraction. Over recent years, a rich literature has emerged concerning the use of ontologies in information extraction, and a number of important tools have been developed to support OBIE [3]. By converting

the cultural model into an ontology using standard knowledge representation languages, such as OWL, we are able to capitalize on the availability of these pre-existing tools and techniques, and we are also able to compare the success of our approach with other OBIE approaches.

The ontologies developed to represent the contents of cultural models are based around the notion of ideas as being divided into concepts, causal beliefs and values. These three types of ideas constitute the top-level constructs of the ontology, and subtypes of these constructs are created to represent the various elements of the cultural model. For example, if we consider the notion of "Jihad support", as depicted in Fig. 1, then we can see that this is a type of concept, and it is regarded as a positive thing, at least from the perspective of the target group. Within the ontology developed to support this cultural model we have the concept of 'support-for-jihad', which is represented as a type of 'jihad-related-action', which is in turn represented as a type of 'action', which is in turn represented as a type of 'concept'. Given the focus of the cultural model in representing causal beliefs, the notions of actions, events and the causally-significant linkages between these types of concept are often the most important elements of the cultural ontology.

3.3 Step 3: Develop OBIE Capability

In this step of the process, the aim is to create rules that automatically detect instances of the concepts, beliefs and values contained in the cultural model. There are clearly a number of ways in which this might be accomplished, especially once one considers the rich array of information extraction techniques and technologies that are currently available [7]; however, we prefer an approach that delivers symbolic extraction rules (i.e. rules that are defined over the linguistic features and lexical elements of the source texts) because the knowledge contained in the rules can be easily edited by subject matter experts. In addition, it is easier to provide explanation-based facilities for rule-based symbolic systems than it is for systems based on statistical techniques.

The approach to rule creation that we have adopted in the context of the IEXTREME project is based on the use of the $(LP)^2$ learning algorithm, which has been used to create a number of semantic annotation systems [4, 5]. The basic approach is to manually annotate a limited number of source texts using the cultural ontology that was created in the previous step. These annotated texts are then used as the training corpus for rule induction (see [8] for more details). During rule induction, the $(LP)^2$ algorithm generalizes from an initial rule that is created from a user-defined example by using generic shallow knowledge about natural language. This knowledge is provided by a variety of NLP resources, such as a morphological analyzer, a part-of-speech (POS) tagger and a gazetteer. The rules that result from the learning process thus incorporate a variety of lexical and linguistic features. Previous research has suggested that rules can be defined over a large number of features. For example, Bontcheva et al [9] used a variety of NLP tools to generate 94 features over which information extraction rules could be defined. Of course, not all of these features are likely to be of equal importance

in creating information extraction systems, and further empirical studies are required to assess their relative value in the domain of cultural modeling.

3.4 Step 4: Extract Concepts

Once extraction rules have been created, they can be applied to potential knowledge sources in order to detect occurrences of the various ideas expressed in the cultural model. Because most of the user-defined annotations will be based on the nodal elements of the cultural model networks, such as those seen in Fig. 1, this step is particularly useful for detecting mentions of specific concepts in source texts. In the case of Sunni extremist cultural models this could, for example, include mentions of jihad-related concepts, for example "Jihad is a means to expel the Western occupiers", as well as references to aspects of Western ideology. In general, information extraction systems based on the machine learning technique described above (i.e. the $(LP)^2$ algorithm) have proved highly effective in identifying instances of the terms defined in an ontology, so we can expect reasonable extraction performance for this step of the process.

3.5 Step 5: Extract Causal Relations

There have been a number of attempts to extract relational information in a Web-based context (see [7]). The use of ontologies in such systems plays an important role because they provide background knowledge about the possible semantic relationships that are likely to exist between the various entities identified in previous processing steps. Thus, if a system first subjects a textual resource to entity-based semantic annotation, then it is able to use the ontology to form expectations about the kind of relationships that might be apparent in particular text fragments. When this background knowledge is combined with lexical and linguistic information, a relation extraction system is often able to identify relationships that would be impossible to detect using a text-only analysis.

The approach to relation extraction that we have adopted in the case of the IEXTREME project is based on a technique that was previously developed to support information extraction in the domain of artists and artistic works [10]. The approach builds on the outcome of the previous step, which is concerned with the detection of concepts in the source texts. Importantly, once these concept annotations are in place, the relation extraction subsystem is provided with a much richer analytic substrate than would otherwise have been the case. In fact, it is only once such annotations are in place that the real value of the ontology (for the detection and extraction of relationships) can be appreciated. In particular, the ontology provides background knowledge that drives the formation of expectations about the kinds of relationships that could appear between concepts, and once such expectations have been established, they can be supported or undermined by subsequent lexical analysis of the sentence in which the concepts occur.

Obviously, the nature of the natural language processing that is performed on the sentence is key to this relation-based annotation capability. It is not sufficient for a system to simply form an expectation about the kind of relationships that might occur between identified entities in the text; the system also needs to ascertain whether the linguistic context of the sentence supports the assertion of a particular relationship. The decision regarding which relationship (if any) to assert in a particular sentential context is based on a strategy similar to that used in previous research [10]. Essentially, each relationship in the ontology is associated with a 'synset' (a set of synonyms) in the general-purpose lexical database WordNet [6]. When the relation extraction system executes, it attempts to match the words in a sentence against the WordNet-based linguistic grounding provided for each expected relationship. In addition to representing information about synonyms, the WordNet database also represents hypernymy (superordinate) and hyponymy (subordinate) relationships. These can be used to support the matching process by avoiding problems due to transliteration.

3.6 Step 6: Exploit Knowledge Extraction Capability

The knowledge extraction capability outlined in the previous steps provides support for the refinement, extension and validation of the knowledge contained in cultural models. The ability to detect instances of the ideas expressed in cultural models across a range of Web resources (including blogs, organizational websites, discussion threads and so on) provides a means by which new knowledge sources can be discovered and made available for a variety of further model development and refinement activities. The use of OBIE technology therefore provides a means by which the latent potential of the Web to serve as a source of culture-relevant knowledge and information can be exploited in the context of qualitative cultural modeling initiatives.

Aside from the development of better qualitative cultural models, the use of knowledge extraction techniques can also support the development of quantitative cultural models. As discussed above, quantitative cultural models extend qualitative cultural models by including information about the relative frequencies of particular ideas within the population to which the cultural model applies [2]. By harnessing the power of OBIE methods, the current approach provides a means by which ideas (most notably concepts and causal beliefs) can be detected across many hundreds, if not thousands, of Web resources. This provides an estimate of the prevalence of particular ideas in the target population of interest, and it provides a means by which the Web can be used to support the development of quantitative cultural models.

4 Conclusion

This paper has described an approach to harnessing the latent potential of the Web to support cultural modeling efforts. The approach is based on the development of culture-oriented knowledge extraction capabilities and the use of

techniques that support a cognitive characterization of specific cultural groups. Systems developed to support the approach may be seen as an important element of iterative cultural modeling efforts, especially ones in which an initial qualitative cultural model drives the acquisition of information from a large number of heterogeneous Web-based resources.

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

References

1. Sperber, D.: Explaining Culture: A Naturalistic Approach. Blackwell, Malden, Massachusetts, USA (1996)
2. Sieck, W.R., Rasmussen, L., Smart, P.R.: Cultural Network Analysis: A Cognitive Approach to Cultural Modeling. In: Verma, D. (ed.) Network Science for Military Coalition Operations: Information Extraction and Interaction. IGI Global, Hershey, Pennsylvania, USA (2010)
3. Wimalasuriya, D.C., Dou, D.: Ontology-based information extraction: An introduction and a survey of current approaches. Journal of Information Science, 36, 3, 306–323 (2010)
4. Vargas-Vera, M., Motta, E., Domingue, J., Lanzoni, M., Stutt, A., Ciravegna, F.: MnM: ontology driven semi-automatic or automatic support for semantic markup. In: 13th International Conference on Knowledge Engineering and Knowledge Management, Siguenza, Spain (2002)
5. Ciravegna, F., Wilks, Y.: Designing adaptive information extraction for the Semantic Web in Amilcare. In: Handschuh, S., Staab, S. (eds.) Annotation for the Semantic Web. IOS Press, Amsterdam, Netherlands (2003)
6. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. J.: Introduction to WordNet: An On-line Lexical Database. International Journal of Lexicography, 3, 4, 235–244 (2004)
7. Sarawagi, S.: Information extraction. Foundations and Trends in Databases, 1, 3, 261–377 (2008)
8. Ciravegna, F.: Adaptive information extraction from text by rule induction and generalisation. In: 17th International Joint Conference on Artificial Intelligence, Seattle, Washington, USA (2001)
9. Bontcheva, K., Davis, B., Funk, A., Li, Y., Wang, T.: Human Language Technologies. In: Davies, J., Grobelnik, M., Mladenic, D. (eds.) Semantic Knowledge Management: Integrating Ontology Management, Knowledge Discovery, and Human Language Technologies. Springer-Verlag, Berlin, Germany (2009)
10. Alani, H., Kim, S., Millard, D. E., Weal, M. J., Hall, W., Lewis, P., Shadbolt, N. R.: Automatic Ontology-Based Knowledge Extraction from Web Documents. IEEE Intelligent Systems, 18, 1, 14–21 (2003)