Tracking and Modeling Information Diffusion across Interactive Online Media

Arno Scharl*
Department of New Media Technology,
MODUL University Vienna, Austria
E-mail: scharl@modul.ac.at
*Corresponding author

Albert Weichselbraun
Department of Applied Computer Science,
Vienna University of Economics and Business Administration, Austria
E-mail: albert.weichselbraun@wu-wien.ac.at

Wei Liu
School of Computer Science and Software Engineering,
University of Western Australia, Perth, Australia
E-mail: wei@csse.uwa.edu.au

Abstract: Information spreads rapidly across Web sites, Web logs and online forums. This paper describes the research framework of the IDIOM Project (Information Diffusion across Interactive Online Media),[1] which analyzes this process by identifying redundant content elements, mapping them to an ontological knowledge structure, and tracking their temporal and geographic distribution. Linguists define "idiom" as an expression whose meaning is different from the literal meanings of its component words. Similarly, the study of information diffusion promises insights that cannot be inferred from individual network elements. This paper presents underlying technology, initial results, and the future roadmap of investigating information diffusion based on ontological knowledge structures. Similar projects often focus on particular media, or neglect important aspects of the human language. This paper addresses these gaps to reveal fundamental mechanisms of information diffusion across media with distinct interactive characteristics.

Keywords: information diffusion, ontology extension, natural language processing.

Reference to this paper should be made as follows: Scharl, A., Weichselbraun, A. and Liu, W. (2007) 'Tracking and Modeling Information Diffusion across Interactive Online Media', *Int. J. Metadata, Semantics and Ontologies*, Vol. 2, No. 3, pp. 43–54.

1 INTRODUCTION

Building upon previous research on media monitoring and Web assessment, IDIOM converts and aggregates data gathered via a Web crawler into annotated content repositories. Understanding electronic content at such an abstract level is crucial in a time of mainstream Internet adoption, when society is looking for new ways to cope with the explosive growth and reduced lifespan of human knowledge (Huberman and Adamic, 1999). The size and increasing complexity of information networks, however, often conceal important trends and correlations. It is possible to track such hidden regularities in order to understand the nature and dynamics of electronic content.

Such understanding will enable organizations to measure and channel information flows. The *Climate Change Collaboratory*, for example, is an initiative of the ECOresearch Network[2] that investigates how knowledge about domain-specific information flows can help achieve a critical mass of participants and content to ensure rich, self-sustaining community interaction.

Knowledge about the mechanisms of information diffusion also helps to retrieve, analyze and distribute information effectively. Organizations can increase the impact of their marketing and public awareness campaigns, and measure this impact accurately. Policy makers gain a detailed understanding of how information replicates within and across interactive environments, and how this process shapes public opinion. Individuals benefit from improved collaboration tools with intuitive visual interfaces to access complex data. Such collaborative systems assist researchers of different disciplines in sharing data and expertise within interdisciplinary environments, thereby contributing to methodological pluralism and catalyzing collective strategies for managing knowledge.

The following Section 2 reviews the literature and states the fundamental research questions underlying this paper. Section 3 then introduces the webLyzard suite of Web mining tools, and describes a number of key components relevant to this research. Sections 3.1-3.5 then present a detailed roadmap to measuring, classifying and visualizing spikes in electronic content including their frequency, intensity and semantic orientation. Section 4 summarizes the research framework and discusses its relevance from an organizational perspective.

2 RESEARCH QUESTIONS

Media richness theory (Daft, Lengel, & Trevino, 1987), Web site usability (Palmer, 2002), competitive intelligence (Hsinchun Chen, Chau, & Zeng, 2002) and service quality (Zeithaml, Parasuraman, & Malhotra, 2003) are common theoretical frameworks to investigate content production. Traditional investigations often rely on individual judgments by experts or survey participants that use lists of weighted attributes (Olsina & Rossi, 2002). Expert evaluations approximate these attributes and thereby introduce varying degrees of subjectivity, while user questionnaires suffer from respondent inaccuracy due to differences between reported and actual behavior. Automatically gathering data from electronic media, by contrast, provide scalability, speed, consistency and abundant longitudinal data. Automated approaches alleviate methodological limitations of subjective impressions and anecdotal evidence (Bauer & Scharl, 2000; Scharl, 2000). Although Web crawlers cannot replace human evaluation, they handle dynamic data more efficiently and help avoid inter- and intra-personal variances.

IDIOM uses a Web crawler to track information diffusion within and across different interactive environments: (i) Web sites of news media, commercial organizations and advocacy groups; (ii) news distribution networks based on the *RDF Site Summary* (RSS) format; (iii) low-overhead forms of personal publishing such as Web logs (“blogs”) and online discussion forums; (iv) communication via electronic mail and instant messaging (since analyzing inter-individual communication raises privacy concerns, transcripts of user interactions often have to be generated through dedicated experiments). Investigating the production, propagation and consumption of content in environments with distinct interactive characteristics, we address four fundamental research questions:

- How widespread is content redundancy in information networks, and what are the factors influencing content replication within and across these networks?
- Does the medium’s degree of interactivity affect information diffusion? And if so, can existing models such as hub-and-spoke, syndication and peer-to-peer explain this influence?
- How does macroscopic information flow shape public opinion? What are appropriate methods to investigate the extent, dynamics and latency of this process?
- What content placement strategies increase the impact on the target audience and support self-reinforcing content propagation in a particular medium?

Answering these questions requires advances in measuring, analyzing and predicting spatial and temporal flows of information. The complexity of the human language, for example, calls for semantic disambiguation (Seo, Chung, Rim, Myaeng, & Kim, 2004) and software components able to *understand* content (Kiryakov, Popov, Terziev, Manov, & Ognyanoff, 2004). Semantic Web technologies as outlined by Berners-Lee (Berners-Lee, Hendler, & Lassili, 2001) and in-depth semantic analyses complement approaches based on graph theory (Dill et al., 2002; Song, Havlin, & Makse, 2005), which represent information networks as a mere set of interconnected nodes.

Gruhl et al. distinguish two types of information diffusion: *spikes* (externally induced sharp rises in activity), and *chatter* (internally driven, sustained discussions). The frequency and shape of spikes is a powerful indicator of information diffusion (Gruhl, Guha, Liben-Nowell, & Tomkins, 2004). Occasionally, spikes result from chatter through a process of resonance, where insignificant exogenous events trigger massive reactions. Such sensitive dependence on initial conditions occurs when large sets of individual interactions generate large-scale, collective behavior. Social network analysis investigates such interactions between people, groups and organizations (Haythornthwaite, 1996; Watts, 2003). By disseminating information via their social networks, individuals create strong peer influence that often surpasses exogenous influences. *Viral marketing* leverages this peer influence to trigger self-reinforcing content propagation among individuals (Godin, 2001; Goldenberg, Libai, & Muller, 2001). Weak and strong ties (Granovetter, 1973) between those individuals determine the distinct paths of information dissemination. It is along these paths that inter-individual communication multiplies the impact of spikes and creates widespread attention.

Revealing the structure and determinants of these paths will guide organizations in their efforts to raise awareness and distribute electronic content. This represents a significant contribution, since integrated projects capturing both spatial and temporal diffusion effects are rare. Previous research on information diffusion neglecting the semantic orientation of electronic content also fails to reflect author attitude, which is an important aspect of the human language (Section 3.2). When analyzing political campaigns (Scharl & Weichselbraun, 2006), for example, the frequency and shape of spikes related to a candidate might prove less significant than the attitude conveyed in these spikes (negative ↔ positive, weak ↔ strong, passive ↔ active, etc.).

Capturing the diversity of the human language through grammatical parsing helps classify electronic content correctly (Section 3.1). Integrating the resulting classifications with external taxonomies yields *seed ontologies* for specific domains, continuously validated, refined and extended via semi-automated methods (Section 3.3). The dynamically updated ontological structures can then be used to contrast conceptual similarity on the document level with textual similarity on the paragraph and sentence level – distinguishing identical copies, reworded segments, and independent articles on the same topic (Section 3.4). This distinction reveals *content redundancy* at a very granular level, and allows tracking the spatial reach and temporal gradient of spikes in electronic content.

Identifying diffusion patterns across millions of network nodes is complex and computationally expensive. A service-oriented architecture will address this issue, utilizing resources effectively and thus allowing significant increases in sample size and measurement frequency (Section 3.5). The service-oriented architecture also provides an interoperable environment enabling global collaboration among researchers with common goals and interests. The dynamic and multi-dimensional nature of information diffusion complicates its analysis and the interpretation of results. This problem can be overcome by using advanced visualization algorithms to increase the transparency of information flows. Such algorithms take advantage of the human ability to recognize visual patterns, and to track movements in two- or three-dimensional graphical environments.

3 METHODOLOGY

This research builds upon webLyzard,[3] a stable and tested platform for media monitoring and large-scale Web assessment. As the volume and dynamic character of Web content entail ongoing analysis, the webLyzard crawling agent mirrors Web sites in monthly or weekly intervals and has amassed over one terabyte of Web data since 1999. In contrast to many Web annotation and analysis projects, webLyzard uses explicitly defined samples of Web sites. The current database of more than 7,000 sites includes, for example, the *Fortune 1000* and the *Fortune Global 500*,[4] more than 150 international news media, and the *Business Review Weekly's* ranking of Australia's 1000 largest corporations.[5]

While processing markup tags and scripting elements, the crawler collects the raw text including headings, menus and link descriptors. Ignoring graphics and multimedia files, the crawler follows a site's hierarchical structure until it reaches a user-specified limit. The current system analyzes 10 megabytes of regular sites and 50 megabytes of news media, but is not limited to these values. While size restrictions facilitate comparative studies and ensure that prominent information is not "diluted" by content of lower hierarchical levels, comprehensive measures of information diffusion will require continuous crawling of the World Wide Web and other media such as e-mail archives or online discussion forums (Section 3.5).

The system architecture of Figure 1 distinguishes proprietary modules, major data structures and embedded third-party tools. The multithreaded *Mirror Control* and *Mirror Storage* modules gather the documents and store them in a compressed archive. This process involves a number of third-party modules to convert PDF, Postscript and word processor files into HTML format.

The *Structural Parser* corrects syntactical errors (e.g. missing elements or misaligned tags) and codes the mirrored data. The resulting site profiles include three groups of variables: (i) *navigational mechanisms:* structure and accessibility of links within and between documents; (ii) *interactive features* such as forms, scripts and Java applets; (iii) *layout and multimedia characteristics* such as frames, tables and embedded images. The *Textual Parser* segments the textual chain into sites, documents and sentences. The hierarchically organized, XML-encoded output file thus preserves the original site structure. The module also calculates linguistic metrics such as the average lengths of content units and the type-token ratio describing the richness of a site's vocabulary. It then automatically detects languages and removes redundant segments that might bias the results. Typical examples of redundant segments *within sites* are non-contextual navigational elements or news headlines appearing on multiple pages (identifying redundant elements *across sites* helps distinguish content creation from content propagation; Section 3.4).

The *Part of Speech Tagging* component integrates suffix analysis with automated learning on pre-tagged corpora to eliminate ambiguities and increase the validity of linguistic metrics. *Word Stemming* addresses syntactic variations that complicate the interpretation of word lists. This optional component puts verb forms into the infinitive, nouns into the singular, and converts elisions. Stemming and frequency thresholds reduce the vocabulary and improve the results' stability (Lebart, Salem, & Berry, 1998). The current system uses a list of English lemmas containing 40,569 words in 14,762 lemma groups (Someya, 1998b), optionally integrating the Porter Stemming Algorithm (Porter, 1980) of the Natural Language Toolkit.[6]

Keyword Analysis locates words in a given text and compares their frequency with a reference distribution from a larger corpus of text. A chi-square test of significance with Yates' correction for continuity determines over-represented terms and lists them in order of decreasing significance. Extending the keyword algorithm, the *Term Co-Occurrence* module uses a pattern matching algorithm based on regular expressions to identify text fragments frequently appearing within the same sentences or documents. When formulating regular expressions, analysts have to enumerate common inflections of a term while excluding general terms with ambiguous meanings.

The *Output and Configuration* layer based on the Zope Application Server[7] and the Plone Content Management System[8] provides the interface to manage samples, update the database and export results for further processing in external applications.

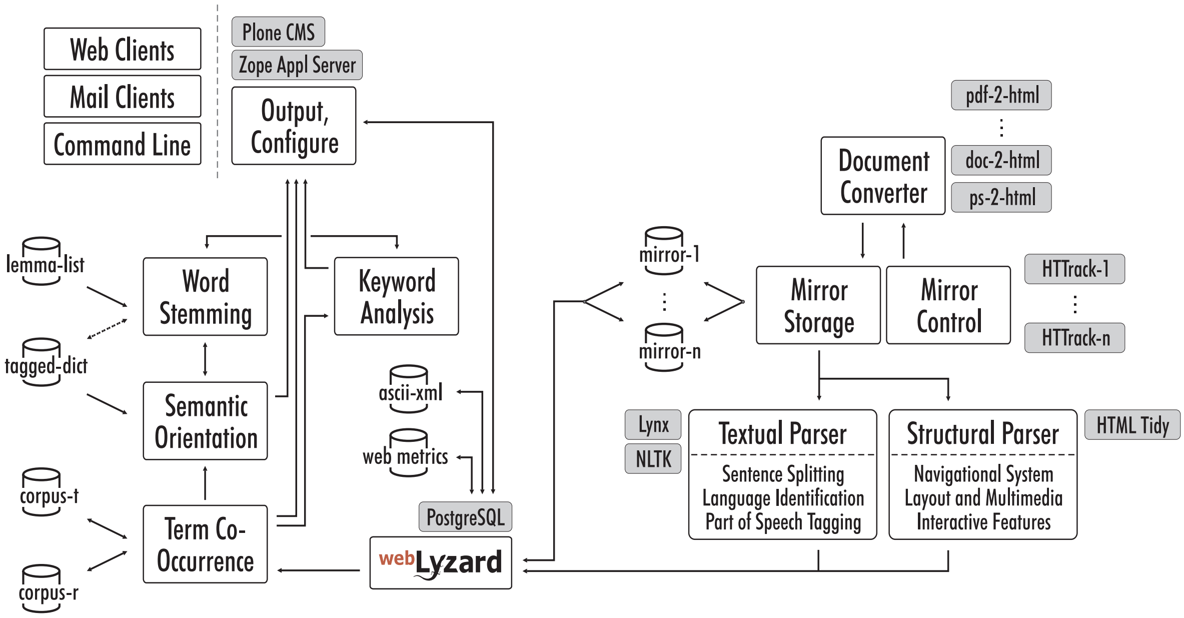

Figure 1 webLyzard system architecture (modules, data structures and third-party tools)

3.1 Tracking and Classifying Information Flows

Since semantic technologies unfold their full potential through network effects, they require a critical mass of annotations (Benjamins et al., 2004; Corcho 2006). Topic classification via semantic annotation, a key element of tracking related information, is no exception. But manual annotation is difficult, time consuming and expensive. Automatic classification attempts to overcome the Web's current lack of semantic annotation. Capturing the diversity of the human language as outlined in Section 3.2 helps to classify electronic content correctly, and to provide a semantic label bureau service (Dill et al., 2003)[9] for participating researchers – using a modified version of the Bayes algorithm (Weichselbraun, 2004), and specifying domain knowledge via formal ontologies (see Section 3.3).

Extensions of the current prototype (Weichselbraun, 2004) will support hierarchical classification (McCallum, Rosenfeld, Mitchell, & Ng, 1998), implement subtopic detection (Hearst, 1997) and refine the prototype's feature selection algorithm (Ahmad & Dey, 2004). Bayesian noise reduction (Zdziarski, 2004), a special case of feature selection, will improve the classifier's accuracy by removing relevant but sparse data. This technique can detect text fragments different from the document's primary classification, and trace these fragments in non-related documents.

3.2 Semantic Orientation of Media Coverage

Research on information diffusion neglecting the semantic orientation of electronic content fails to reflect author attitude (e.g. positive versus negative), which is an important aspect of the human language. The lack of local context also limits the explanatory power of word frequency data (Biber, Conrad, & Reppen, 1998; McEnery & Wilson, 1996). Assuming that text segments reflect local coherence, author attitude towards specific topics can be inferred from the distance between a target term and sentiment words from a tagged dictionary (Scharl, Pollach, & Bauer, 2003).

The current dictionary uses 4,400 positive and negative sentiment words from the General Inquirer (Stone, 1997). Reverse lemmatization added about 3,000 terms to the dictionary by considering plurals, gerund forms, past tense suffixes and other syntactical variations (e.g. MANIPULATE → MANIPULATES, MANIPULATING, MANIPULATED). Adding multiple-word combinations to the tagged dictionary to discern morphologically similar but semantically different terms such as FUEL CELL and PRISON CELL should further increase the method's accuracy. Yet the lexis of electronic content only partially determines its semantic orientation, despite using multi-word units of meaning (Danielsson, 2004) instead of single words or lemmas. In its later phases, the IDIOM project will employ grammatical parsing to address this limitation, resolving ambiguities and capture meaning-making processes at levels beyond lexis.

Words with different or even opposite meanings, depending on the context, represent an inherent problem of automatically determining semantic orientation. ARREST as a noun takes custody by legal authority, for example, while ARREST as a verb means to catch or stop. Similarly, the adjective GOOD assigns desirable or positive qualities. In economics, however, the noun GOOD refers to physical objects or services. Part of speech tagging considers this variability by annotating terms and distinguishing grammatical categories such as article (AT), noun (NN), verb (VB), adverb (RB), past-tense-verb (VBD), object pronoun (PPO) and possessive pronoun (PP$). The sentence *"He still saw her",* for example, would be annotated with PPO RB VBD PPO. Heterogeneous language use complicates this annotation process, but also represents an opportunity to identify cultural determinants of content production. Term frequency, spelling and usage context differ across media; instant messaging acronyms, for example, rarely appear on corporate Web sites.

3.3 Validating and Extending Ontological Structures

One of the motivations for building ontologies is establishing shared meaning via a commonly agreed terminology (Fensel, Wahlster, Lieberman, & Hendler, 2003; Maedche, 2002). By providing a formal and non-ambiguous terminology in a given domain, ontologies determine the context for classifying information flows. Data extraction and evaluation services integrate ontology knowledge to disambiguate content (Dill et al., 2003; Kiryakov, Popov, Terziev, Manov, & Ognyanoff, 2004), and to track the rise and decay of topics. *OWL Lite*,[10] a sublanguage of the *Ontology Web Language (OWL),* provides a quick migration path for thesauri and other taxonomies. OWL Lite suffices for a context-dependent assessment of tokens, sentences and documents. Incorporating and extending external knowledge repositories improves topic detection and tracking. Potential sources include generic directories such as *TAP* (Guha & McCool, 2003) and *DMOZ*,[11] and domain-specific thesauri such as the *EPA Terminology Reference System*[12] and the *EEA Multilingual Environmental Glossary*.[13]

Given online media's dynamic character, ontologies should be continuously validated and updated through revision, merging and semi-automated extension processes based on aggregated content. *Revision* absorbs single topics or single associations at a time. *Merging* incorporates previously generated ontologies containing multiple topics and associations – equivalent to repeated revisions when the order of incoming topics or associations does not affect their epistemic importance. Validation and semi-automated ontology extension leverage the keyword module by identifying terms related to formal ontology concepts (Feng, Chang, & Dillon, 2002). Applying keyword analysis across hierarchical layers identifies hypo- and hypernyms (words more specific/generic than a given word), computes the "keyness" of terms, and incorporates term verification mechanisms based on external directories such as *Merriam Webster*,[14] *WordNet* (Seo, Chung, Rim, Myaeng, & Kim, 2004)[15] and *OpenCyc*.[16] Interfaces to popular ontology editors like *Protégé* (Noy et al., 2001), *OntoTrack* (Liebig & Noppens, 2005) and *Swoop* (Kalyanpur, Parsia, & Hendler, 2005) will help incorporate suggested changes.

A pilot study identified concept hierarchies using spreading activation on weighted graphs (Liu, Weichselbraun, Scharl, & Chang, 2005). Figure 2 outlines the prototype's system architecture, which answers calls for research into combining different learning paradigms for identifying taxonomic relations from large corpora (Cimiano, Pivk, Schmidt-Thieme, & Staab, 2005). The prototype selects a small set of terms from domain experts or from known ontology repositories as seed ontology, which is then fed into the *Lexical Analyzer*.

Plurals, gerund forms, and past tense suffixes are syntactical variations that complicate the automatic processing of textual information. Lemmatizing the media corpus addresses this problem, putting verb forms into the infinitive, nouns into the singular, and removing elisions. The prototype uses an adapted version of Someya's lemma list containing 40,569 words in 14,762 lemma groups (Someya, 1998a). Lemmatizing the underlying corpus improves the extended ontology's stability and generalizability by clustering terms of similar meaning.

Co-occurrence analysis at both the sentence and the document level limits the influence of popular terms that are not related to the domain (Roussinov & Zhao, 2003). Terms are selected according to a threshold value on the co-occurrence significance. Lexical analysis is done by consulting the WordNet lexical dictionary (Fellbaum, 1998), and by analyzing the Web corpus for terms connected by *trigger phrases*. A trigger phrase matches a fragment of text that contains a parent-child description (Joho, Sanderson, & Beaulieu, 2004), similar to the matching of lexico-syntactic patterns to identify hyponym relations known as *Hearst Patterns* (Hearst, 1992).

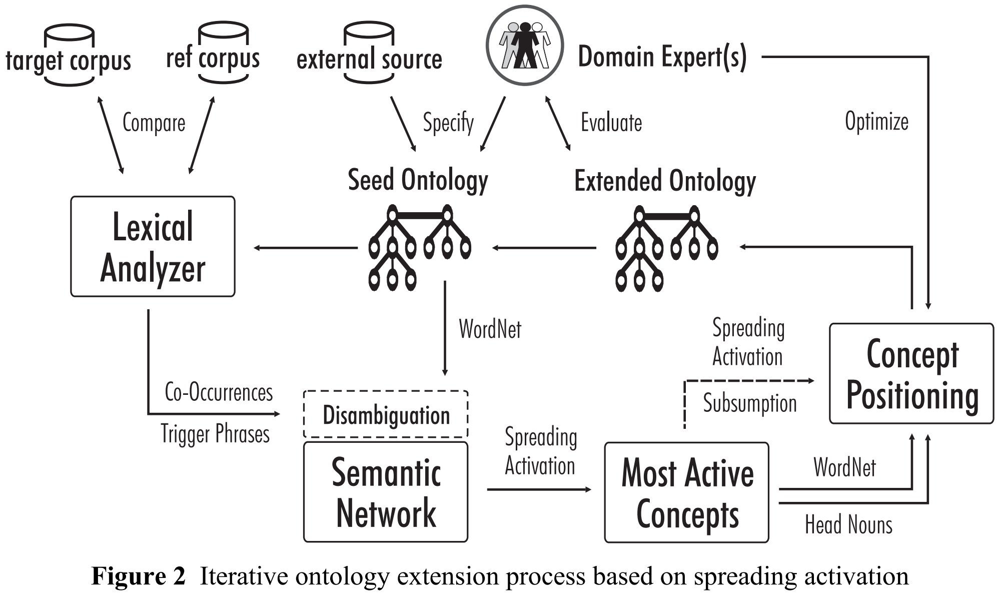

Figure 2 Iterative ontology extension process based on spreading activation

The generated terms are connected with the seed ontology via directed weighted links. Once the network is established, spreading activation identifies the terms most relevant within the domain and suggests their incorporation into the seed ontology. WordNet, head nouns and subsumption analysis are then used to confirm the semantic relationship – considering the relation of components in multi-word terms (Navigli & Velardi, 2004), and the fact that head nouns – e.g. EXTRACTION in the term '*crude oil extraction'* – often super-ordinate the containing phrase. Similarly, common short phrases (SUCH AS, AND OTHER, or INCLUDING) often indicate subsumption.

A machine learning component will extend the iterative loop depicted in Figure 2, optimizing the parameters of the ontology extension process based on feedback from domain experts. Using a portion of a manually validated semantic structure as seed ontology, this adaptive component could also modify the parameters based on the similarity between the validated structure and the extended ontology suggested by the system. This method optimizes results within a given domain. Applying this method to several domains should allow formulating general strategies to extend ontologies with limited domain knowledge, or when facing sources containing conflicting evidence (Cimiano, Pivk, Schmidt-Thieme, & Staab, 2005).

The seed ontology on *energy sources* used for the following example comprises seven concepts: 1. energy sources; 1.1 fossil fuels; 1.1.1 crude oil; 1.1.2 coal; 1.2 renewable energy; 1.2.1 wind energy; 1.2.2 solar energy. Combining the methods outlined above to analyze media coverage of this domain yielded a semantic network of more than a thousand nodes connected via annotated links – link types include *co-occurrence*, *trigger phrase {hyponym, hypernym, synonym}*, *wordnet {hyponym, hypernym, synonym}*, *co-occurrence significance*, *hypernym of the original hierarchy,* and *head noun*.

Hierarchically positioning the activated terms (that is, those most relevant to the domain and seed ontology), represents a challenging task. The current prototype follows three steps: (i) accept semantic relations confirmed by WordNet and head noun analysis; (ii) remove modifiers of a noun phrase that also appear in the activated list, as they do not represent the term's core meaning; (iii) trigger another round of spreading activation using the non-confirmed terms as seed terms to identify appropriate nodes for attaching these terms; use subsumption analysis to determine the type of relationship. Figure 3 shows the extended ontology after two iterations of spreading activation. Arrows indicate hierarchical relationships. Dotted lines represent semantic associations whose hierarchy could not be determined, with the values in brackets indicating the degree of association.

Feedback from domain experts on the automatically generated ontology will help optimize the parameters of the ontology extension algorithm via machine learning algorithms. Incorporating latent semantic analysis, as described in the following section, for both term elicitation and concept positioning will also enhance ontology creation and knowledge acquisition by capturing implicit relations between concepts that never co-occur within the same sentences or documents.

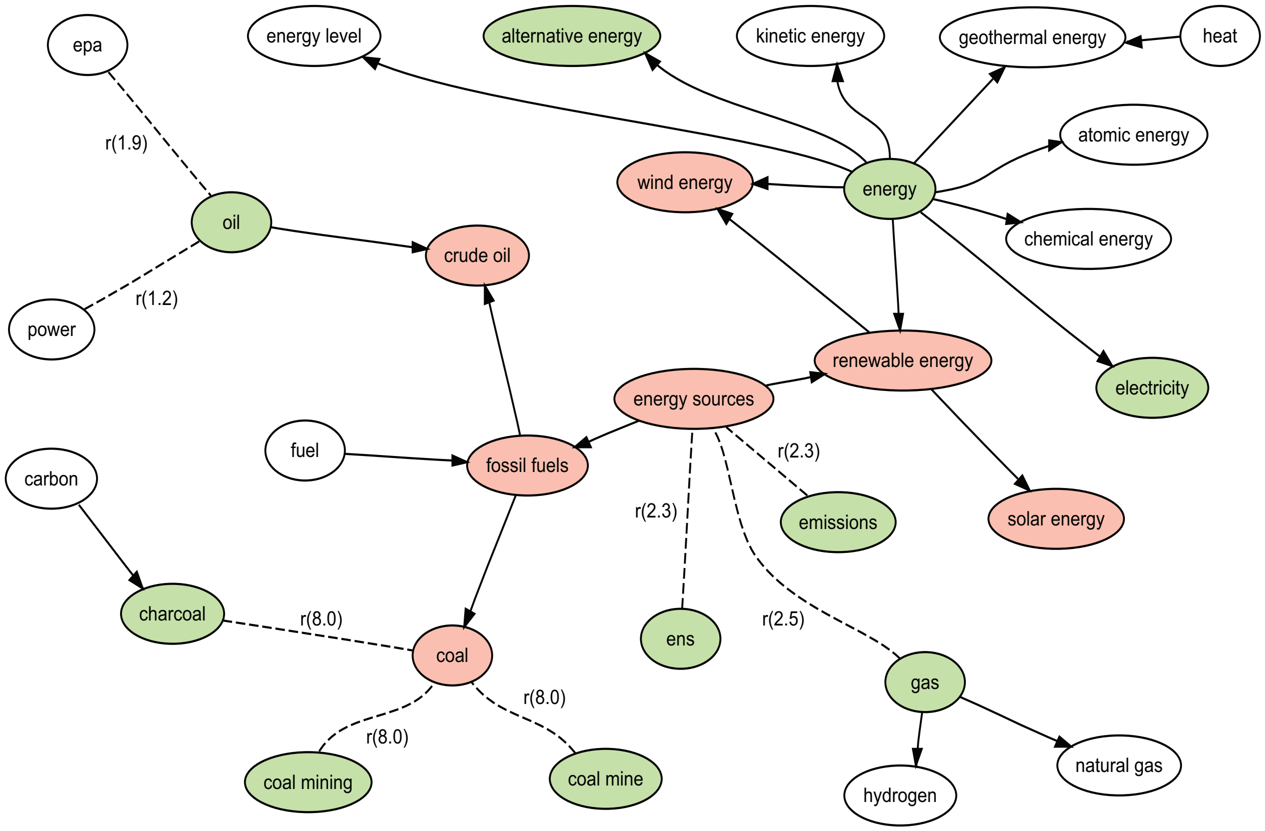

Figure 3 Concept hierarchy after two rounds of spreading activation

3.4 Measuring Content Redundancy

Without a proper analytical framework, content fragments often seem randomly scattered across networks. Tracking and annotating electronic documents to identify identical or similar content fragments should reveal fundamental mechanisms of information diffusion. Popular similarity measures use the classic *vector space model* (Salton, 1989), operating on vector representations that neglect ontological relationships (Bernstein, Kaufmann, Bürki, & Klein, 2005). These methods fail to detect similar meaning in texts with different vocabularies. This restriction led to *latent semantic analysis* (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990) and *concept space approaches* (H. Chen, Hsu, Orwig, Hoopes, & Nunamaker, 1994; Loh, Wives, & de Oliveira, 2000), which provide abstract similarity measures for conceptual comparisons.

After identifying textual segments related to previously unidentified events, IDIOM will employ similarity measures on three levels: document, paragraph and sentence. It will detect languages and classify content segments by ontology concept. The resulting, abstract classifications will be compared with the results of locality-sensitive hashes (Charikar, 2002; Gionis, Indyk, & Motwani, 1999), a granular approach towards textual similarity popular among developers of anti-spam software.[17] This two-fold approach, conceptual similarity on the document level versus textual similarity on the paragraph or sentence level, will distinguish identical copies from reworded segments and independent articles covering the same topic.

3.5 Service-oriented Architecture

The method described in Section 3.4 investigates spatial and temporal information diffusion, which requires a powerful and scalable infrastructure that is capable of handling very large samples to measure spatial effects, with a monitoring frequency high enough to account for temporal effects. Distributed, service-oriented system architectures help meeting these demands. They accelerate data gathering and allow sampling millions of documents by leveraging bandwidth and computational capacity of geographically dispersed systems. Distributed server clusters provide the capacity to include low-overhead forms of personal publishing (Web logs, discussion forums) and transcripts of inter-individual communication (electronic mail, instant messaging). Adding such heterogeneous sources and replacing discrete mirroring intervals by continuous network crawling poses new challenges to sample management, and require accurate methods of determining a node's position in the virtual space.

There are three approaches to distributed information processing (Scharl, 2004): *Peer-to-Peer (P2P) computing* targets applications with a high ratio of computation to data; otherwise gains might be offset by bandwidth overheads. *Grid computing* serves moderate-sized communities and emphasizes resource integration in environments of at least limited trust (Foster & Iamnitchi, 2003). *Web services* support the Web's evolution from a document repository to a service-oriented infrastructure coordinating distributed resources. While IDIOM would benefit from the scalability and fault tolerance of P2P computing, its data-intensity suggests a light-weight Grid strategy as pursued by the *Knowledge Grid* (Cannataro & Talia, 2003), a knowledge extraction service on top of the *Globus Toolkit*.[18] Migrating to such a distributed architecture is complex and labor-intensive. Therefore, after formally evaluating available options in cooperation with the *Australian Partnership for Advanced Computing*,[19] IDIOM will use a standard service layer to provide the core Grid functionality. This will simplify system implementation and maintenance, help manage computing resources effectively, and facilitate collaboration with other research teams investigating and developing service-oriented architectures. To encourage a collaborative development of analytical modules, the Grid-enabled service layer will provide remote access to computational resources – i.e., offering raw and aggregated data in a variety of formats, and facilitating access to the underlying document repository.

4 CONCLUSION

Media monitoring provides a unique empirical base to unveil the conditions that lead to the introduction, transfer and uptake of knowledge. By detecting regularities, agglomerating content, pinpointing trends and determining success factors of networked information systems, automated systems allow identifying and exploiting the potential of new media for knowledge discovery and knowledge management. The presented method to extend and validate ontological structures automatically hints at the potential of such a comprehensive media monitoring framework.

Multiple disciplines need to be integrated in order to analyze the determinants, structure and impact of electronic content. Such an analysis also requires in-depth knowledge about the domain under investigation. The IDIOM use cases specifically address travel and tourism, sustainability, and political communication. The *US Election 2008 Web Monitor,*[20] for example, will provide the empirical means to assess the role and impact of electronic media in political campaigns, and to investigate how macroscopic information flows shape public opinion.

A better understanding of notoriously volatile and often redundant electronic content will create opportunities for organizations and individuals alike. Companies learn how to multiply the impact of their marketing campaigns, and how to accurately measure this impact. Policy makers gain a detailed understanding of how information replicates in interactive environments, and how this process influences public opinion. Such in-depth knowledge about the structure and determinants of information diffusion in online media will guide organizations in placing electronic content to generate viral marketing effects and trigger self-reinforcing content propagation among individuals.

ACKNOWLEDGEMENT

The IDIOM Project is funded by the Austrian Ministry of Transport, Innovation & Technology and the Austrian Research Promotion Agency within the strategic objective FIT-IT (www.fit-it.at).

REFERENCES

Ahmad, A., & Dey, L. (2004). A Feature Selection Technique for Classificatory Analysis. *Pattern Recognition Letters, 26*(1), 43-56.

Bauer, C., & Scharl, A. (2000). Quantitative Evaluation of Web Site Content and Structure. *Internet Research, 10*(1), 31-43.

Benjamins, R., Contreras, J., Corcho, O., & Gómez-Pérez, A. (2004). Six Challenges for the Semantic Web. *AIS SIGSEMIS Bulletin, 1*(1), 24-25.

Berners-Lee, T., Hendler, J., & Lassili, O. (2001). The Semantic Web. *Scientific American, 284*(5), 28-37.

Bernstein, A., Kaufmann, E., Bürki, C., & Klein, M. (2005). *How Similar Is It? Towards Personalized Similarity Measures in Ontologies.* Paper presented at the 7. International Tagung Wirtschaftsinformatik, Bamberg, Germany.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics - Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Cannataro, M., & Talia, D. (2003). The Knowledge Grid. *Communications of the ACM, 46*(1), 89-93.

Charikar, M. S. (2002). *Similarity Estimation Techniques from Rounding Algorithms.* Paper presented at the 34th Annual ACM Symposium on Theory of Computing, Montreal, Canada.

Chen, H., Chau, M., & Zeng, D. (2002). CI Spider: A Tool for Competitive Intelligence on the Web. *Decision Support Systems, 34*(1), 1-17.

Chen, H., Hsu, P., Orwig, R., Hoopes, L., & Nunamaker, J. F. (1994). Automatic Concept Classification of Text from Electronic Meetings. *Communications of the ACM, 37*(10), 56-73.

Cimiano, P., Pivk, A., Schmidt-Thieme, L., & Staab, S. (2005). Learning Taxonomic Relations From Heterogeneous Evidence. In P. Buitelaar, P. Cimiano & B. Magnini (Eds.), *Ontology Learning from Text: Methods, Evaluation and Applications* (pp. 59-73). Amsterdam: IOS Press.

Corcho, O. (2006). Ontology-based Document Annotation: Trends and Open Research Problems, *International Journal of Metadata, Semantics and Ontologies,* 1(1), 47-57.

Daft, R. L., Lengel, R. H., & Trevino, L. K. (1987). Message Equivocality, Media Selection, and Manager Performance: Implications for Information Systems. *MIS Quarterly,* 11(3), 355-366.

Danielsson, P. (2004). Automatic Extraction of Meaningful Units from Corpora. *International Journal of Corpus Linguistics, 8*(1), 109-127.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science, 41*(6), 391-407.

Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., et al. (2003). A Case for Automated Large-Scale Semantic Annotation. *Journal of Web Semantics, 1*(1), 115-132.

Dill, S., Kumar, R., McCurley, K. S., Rajagopalan, S., Sivakumar, D., & Tomkins, A. (2002). Self-Similarity in the Web. *ACM Transactions on Internet Technology, 2*(3), 205-223.

Fellbaum, C. (1998). WordNet An Electronic Lexical Database. *Computational Linguistics, 25*(2), 292-296.

Feng, L., Chang, E., & Dillon, T. (2002). A Semantic Network-based Design Methodology for XML Documents. *ACM Transactions on Information Systems, 20*(4), 390-421.

Fensel, D., Wahlster, W., Lieberman, H., & Hendler, J. (Eds.). (2003). *Spinning the Semantic Web - Bringing the World Wide Web to Its Full Potential*. Cambridge: MIT Press.

Foster, I., & Iamnitchi, A. (2003). On Death, Taxes, and the Convergence of Peer-to-Peer and Grid Computing. In F. Kaashoek & I. Stoica (Eds.), *Peer-to-Peer Systems II: Second International Workshop, IPTPS 2003 Berkeley, CA, USA (Lecture Notes in Computer Science, Vol 2735)* (pp. 118-128). Heidelberg: Springer.

Gionis, A., Indyk, P., & Motwani, R. (1999). *Similarity Search in High Dimensions via Hashing.* Paper presented at the 25th International Conference on Very Large Data Bases, Edinburgh, UK.

Godin, S. (2001). *Unleashing the Idea Virus*. New York: Hyperion.

Goldenberg, J., Libai, B., & Muller, E. (2001). Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth. *Marketing Letters, 12*(3), 209-221.

Granovetter, M. (1973). The Strength of Weak Ties. *American Journal of Sociology, 78*(6), 1360-1380.

Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2004). *Information Diffusion Through Blogspace.* Paper presented at the 13th International World Wide Web Conference, New York, USA.

Guha, R. V., & McCool, R. (2003). TAP: A Semantic Web Platform. *Computer Networks, 42*(5), 557-577.

Haythornthwaite, C. (1996). Social Network Analysis: An Approach and Technique for the Study of Information Exchange. *Library & Information Science Research, 18*(4), 323-342.

Hearst, M. A. (1992). *Automatic Acquisition of Hyponyms from Large Text Corpora.* Paper presented at the 14th International Conference on Computational Linguistics, Nantes, France.

Hearst, M. A. (1997). TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages. *Computational Linguistics, 23*(1), 33-64.

Huberman, B.A. & Adamic, L.A. (1999). Growth Dynamics of the World-Wide Web. *Nature,* 401, 131.

Joho, H., Sanderson, M., & Beaulieu, M. (2004). *A Study of User Interaction with a Concept-based Interactive Query Expansion Support Tool.* Paper presented at the Advances in Information Retrieval, 26th European Conference on Information Retrieval.

Kalyanpur, A., Parsia, B., & Hendler, J. (2005). A Tool for Working with Web Ontologies. *International Journal on Semantic Web and Information Systems*, 1(1), 36-49.

Kiryakov, A., Popov, B., Terziev, I., Manov, D., & Ognyanoff, D. (2004). Semantic Annotation, Indexing, and Retrieval. *Web Semantics, 2*(1), 49-79.

Lebart, L., Salem, A., & Berry, L. (1998). *Exploring Textual Data* (Vol. 4). Dordrecht: Kluwer Academic Publishers.

Liebig, T., & Noppens, O. (2005). OntoTrack: A Semantic Approach for Ontology Authoring. *Journal of Web Semantics*, 3(2), 116-131.

Liu, W., Weichselbraun, A., Scharl, A., & Chang, E. (2005). Semi-Automatic Ontology Extension Using Spreading Activation. *Journal of Universal Knowledge Management, 0*(1), 50-58.

Loh, S., Wives, L. K., & de Oliveira, J. P. M. (2000). Concept-based Knowledge Discovery in Texts Extracted from the Web. *ACM SIGKDD Explorations Newsletter, 2*(1), 29-39.

Maedche, A. (2002). *Ontology Learning for the Semantic Web*. Boston: Kluwer Academic.

McCallum, A. K., Rosenfeld, R., Mitchell, T. M., & Ng, A. Y. (1998). Improving Text Classification By Shrinkage in a Hierarchy of Classes. In J. W. Shavlik (Ed.), *15th International Conference on Machine Learning (ICML-98)* (pp. 359-367). Madison, USA: Morgan Kaufmann Publishers.

McEnery, T., & Wilson, A. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Navigli, R., & Velardi, P. (2004). Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites. *Computational Linguistics, 30*(2), 151-179.

Noy, N. F., Sintek, M., Decker, S., Crubezy, M., Fergerson, R. W., & Musen, M. A. (2001). Creating Semantic Web Contents with Protégé-2000. *IEEE Intelligent Systems, 16*(2), 60-71.

Olsina, L., & Rossi, G. (2002). Measuring Web Application Quality with WebQEM. *IEEE Multimedia, 9*(4), 20-29.

Palmer, J. W. (2002). Web Site Usability, Design, and Performance Metrics. *Information Systems Research, 13*(2), 151-167.

Porter, M. (1980). An Algorithm for Suffix Stripping. *Program, 14*(3), 130-137.

Roussinov, D., & Zhao, J. L. (2003). Automatic Discovery of Similarity Relationships through Web Mining. *Decision Support Systems, 35*, 149-166.

Salton, G. (1989). *Automatic Text Processing*. Reading: Addison-Wesley.

Scharl, A. (2000). *Evolutionary Web Development*. London: Springer. http://webdev.wu-wien.ac.at/.

Scharl, A. (2004). A Roadmap Towards Distributed Web Assessment. In N. Koch, P. Fraternali & M. Wirsing (Eds.), *Web Engineering - 4th International Conference, ICWE 2004, Munich, Germany (Lecture Notes in Computer Science, Vol 3140)* (pp. 171-175). Berlin: Springer.

Scharl, A., Pollach, I., & Bauer, C. (2003). Determining the Semantic Orientation of Web-based Corpora. In J. Liu, Y. Cheung & H. Yin (Eds.), *Intelligent Data Engineering and Automated Learning, 4th International Conference, IDEAL-2003, Hong Kong (Lecture Notes in Computer Science, Vol. 2690)* (pp. 840-849). Berlin: Springer.

Scharl, A., & Weichselbraun, A. (2006). *Web Coverage of the 2004 US Presidential Election.* Paper presented at the 2nd International Web as Corpus Workshop, 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006), Trento, Italy.

Seo, H.-C., Chung, H., Rim, H.-C., Myaeng, S. H., & Kim, S.-H. (2004). Unsupervised Word Sense Disambiguation Using WordNet Relatives. *Computer Speech & Language, 18*(3), 253-273.

Someya, Y. (1998a, 01-09-98). e_lemma.txt. Retrieved 20-01-99, 1999, from http://www.lexically.net/downloads/e_lemma.zip

Someya, Y. (1998b). English Lemma List.

Song, C., Havlin, S., & Makse, H. A. (2005). Self-Similarity of Complex Networks. *Nature, 433*(7024), 392-395.

Stone, P. J. (1997). Thematic Text Analysis: New Agendas for Analyzing Text Content. In C. Roberts (Ed.), *Text Analysis for the Social Sciences* (pp. 35-54). Mahwah: Lawrence Erlbaum.

Watts, D. J. (2003). *Six Degrees: The Science of a Connected Age*. New York: W. W. Norton.

Weichselbraun, A. (2004). *Ontology-based Text Classification via Mathematical Methods (in German)*. Unpublished PhD Thesis, Vienna University of Economics and Business Administration.

Zdziarski, J. A. (2004). Bayesian Noise Reduction: Progressive Noise Logic for Statistical Language Analysis. from http://www.nuclearelephant.com/projects/dspam/bnr.html

Zeithaml, V. A., Parasuraman, A., & Malhotra, A. (2003). Service Quality Delivery Through Web Sites: A Critical Review of Extant Knowledge. *Journal of the Academy of Marketing Science, 30*(4), 362-375.

WEBSITES

[1] IDIOM Research Project, http://www.idiom.at/

[2] ECOresearch Network, http://www.ecoresearch.net/

[3] webLyzard, http://www.weblyzard.com/

[4] Fortune Magazine, http://www.fortune.com/

[5] BRW 1000, http://www.brw.com.au/

[6] Natural Language Toolkit, http://nltk.sourceforge.net/

[7] Zope Application Server, http://www.zope.org/

[8] Plone Content Management System, http://www.plone.org/

[9] Platform for Internet Content Selection, http://www.w3.org/PICS

[10] Web Ontology Language, http://www.w3.org/TR/owl-features

[11] DMOZ Open Directory Project, http://dmoz.org/

[12] Terminology Reference System, http://www.epa.gov/trs

[13] Multilingual Environmental Glossary, http://glossary.eea.eu.int/

[14] Merriam-Webster Online, http://www.m-w.com/

[15] WordNet, http://www.cogsci.princeton.edu/~wn

[16] OpenCyc, http://www.opencyc.org/

[17] Apache SpamAssassin, http://spamassassin.apache.org/

[18] Globus Alliance, http://www.globus.org/

[19] APAC, http://www.apac.edu.au/

[20] US Election 2008 Web Monitor, http://www.ecoresearch.net/election2008