

# An Ontology-based Architecture for Tracking Information across Interactive Electronic Environments

Arno Scharl

*Know-Center, Institute of Knowledge Management  
Graz University of Technology  
scharl@ecoresearch.net*

Albert Weichselbraun

*Department of Information Systems and Operations  
Vienna University of Economics and Business Administration  
weichselbraun@ecoresearch.net*

Wei Liu

*School of Computer Science and Software Engineering  
The University of Western Australia  
wei@csse.uwa.edu.au*

## Abstract

*This paper presents technical foundation, roadmap and initial results of the IDIOM project (Information Diffusion across Interactive Online Media). Information spreads rapidly across Web sites, Web logs and online forums. IDIOM tracks this process and compares it to direct communication through electronic mail and instant messaging. 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. Similar projects often focus on particular media, or neglect important aspects of the human language. IDIOM addresses these gaps to reveal fundamental mechanisms of information diffusion across media with distinct interactive characteristics.*

## 1. Introduction

Understanding electronic content at the macro level is crucial in a time of mainstream Internet adoption. Society has to find new ways to cope with the explosive growth and dwindling lifespan of human knowledge. The size and increasing complexity of information networks, however, often conceal important trends and correlations. The IDIOM project aims to track such hidden regularities to understand the nature and dynamics of electronic content. Such understanding will enable policy makers and organizations to measure and channel information flows. Building upon previous research on media monitoring and Web assessment, IDIOM converts data gathered via a Web crawler into annotated content repositories.

Knowledge about the mechanisms of information diffusion will help retrieve, analyze and distribute information effectively. IDIOM will help organizations increase the impact of their marketing and public awareness campaigns, and measure this impact accurately. Policy makers will gain a detailed understanding of how information replicates within and across interactive environments, and how this process shapes public opinion. Individuals will benefit from improved collaboration tools with intuitive visual interfaces to access complex data. The tools provided by IDIOM will encourage researchers of different disciplines to share data and expertise. By providing an interdisciplinary environment, IDIOM will contribute to methodological pluralism and catalyze collectively agreed strategies for distributing information effectively.

## 2. Goals and Significance

Media richness theory [11], Web site usability [40], competitive intelligence [9] and service quality [53] 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 [39]. 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. Crawler-based methods, by contrast, provide scalability, speed, consistency and abundant longitudinal data. They alleviate methodological limitations of subjective impressions and anecdotal evidence [2; 44]. Crawlers cannot replace human eval-

uation, but handle dynamic data more efficiently and help avoid inter- and intra-personal variances.

IDIOM will use crawling agents 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, IDIOM will use publicly available archives and customize the *Generic Remote Usage Monitoring Production System* [16] to generate transcripts of user interactions via dedicated experiments (hosted by the University of Glasgow, this system automatically collects, manages and analyzes large collections of user actions from geographically remote clients).

By investigating the production, propagation and consumption of content in environments with distinct interactive characteristics, IDIOM will 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 [47] and software components able to *understand* content [30]. This is the fundamental idea behind Berners-Lee’s vision of the *Semantic Web* [4]. The in-depth semantic analysis of IDIOM complements approaches based on graph theory [15; 49], which represent information networks as a mere set of interconnected nodes.

IDIOM distinguishes 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 [25]. 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 [27; 51]. 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 [22; 23]. Weak and strong ties [24] 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.

IDIOM will reveal the structure and determinants of these paths to guide organizations in their efforts to raise awareness and distribute electronic content. This is a significant contribution, because 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 3.1). When analyzing political campaigns, 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 will help IDIOM to classify electronic content correctly (Section 3.2). By integrating the resulting classifications with external taxonomies, IDIOM will build *seed ontologies* for specific domains. Semi-automated methods will continuously validate, refine and extend these ontologies (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 lightweight Grid architecture will address this issue, utilizing resources effectively and thus allowing significant increases in sample size and measurement frequency (Section 3.5). The Grid architecture also provides an interoperable environment enabling global collaboration among researchers of similar interests.

The dynamic and multi-dimensional nature of information diffusion complicates its analysis and the interpretation of results. IDIOM will overcome this problem by using advanced visualization algorithms to increase the transparency of information flows. This will take advantage of the human ability to recognize visual patterns, and to track movements in two- or three-dimensional graphical environments.

The following Section 3 outlines the methodology and current system architecture. 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.

### 3. Methodology

IDIOM builds upon webLyzard [69], 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 information retrieval and Web annotation projects, webLyzard analyzes 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* [60], more than 150 international news media, and the *Business Review Weekly's* ranking of Australia's 1000 largest corporations [55].

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.

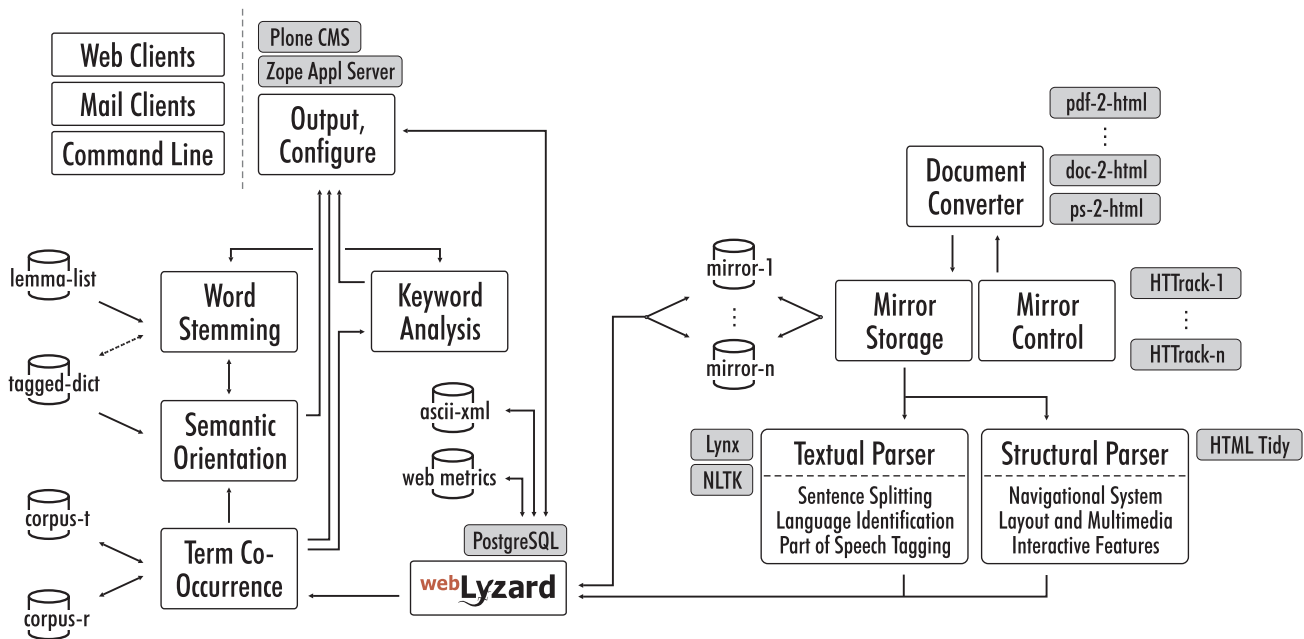


Figure 1. System architecture of the current prototype – modules, data structures and third-party tools

*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 [31]. The current system combines the Porter Stemming Algorithm [41] of the Natural Language Toolkit [63] with a list of English lemmas containing 40,569 words in 14,762 lemma groups [48].

*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 [71] and the Plone Content Management System [66] provides the interface to manage samples, update the database and export results for further processing in external applications.

### 3.1. Semantic Orientation

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 [6; 36]. 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 taken from a tagged dictionary [46].

The current dictionary uses 4,400 positive and negative sentiment words from the General Inquirer [50]. 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).

IDIOM will add multiple-word combinations to the tagged dictionary to discern morphologically similar but semantically different terms such as FUEL CELL and PRISON CELL. Yet the lexis of electronic content only partially determines its semantic orientation, despite using multi-word units of meaning [12] instead of single words or lemmas. IDIOM 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 factors that determine content production. Term frequency, spelling and usage context differ across media; common instant messaging acronyms, for example, rarely appear on corporate Web sites.

### 3.2. Topic Classification

Since semantic technologies unfold their full potential through network effects, they require a critical mass of annotations [3]. 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 3.1 will help IDIOM to classify electronic content correctly, and to provide a semantic label bureau service [14; 68] for participating researchers. This service will use a modified version of the Bayes algorithm and specify domain knowledge via formal ontologies (Section 3.3).

Extensions of the current prototype [32] will support hierarchical classification [35], implement subtopic detection [28] and refine the prototype's feature selection algorithm [1]. Bayesian noise reduction [52], a special case of feature selection, will improve the classifier's accuracy by removing relevant but sparse data. This technique will help detect text fragments different from the document's primary classification, and trace these fragments in non-related documents.

### 3.3. Ontology Validation and Refinement

One of the main motivations for building ontologies [19; 34] is creating shared and commonly agreed meaning for automated knowledge processing. By providing formal, common and non-ambiguous terminology in a given domain, ontologies establish the context for classifying information. IDIOM data extraction and evaluation services integrate ontology knowledge to disambiguate content [14; 30], and to track the rise and decay of topics.

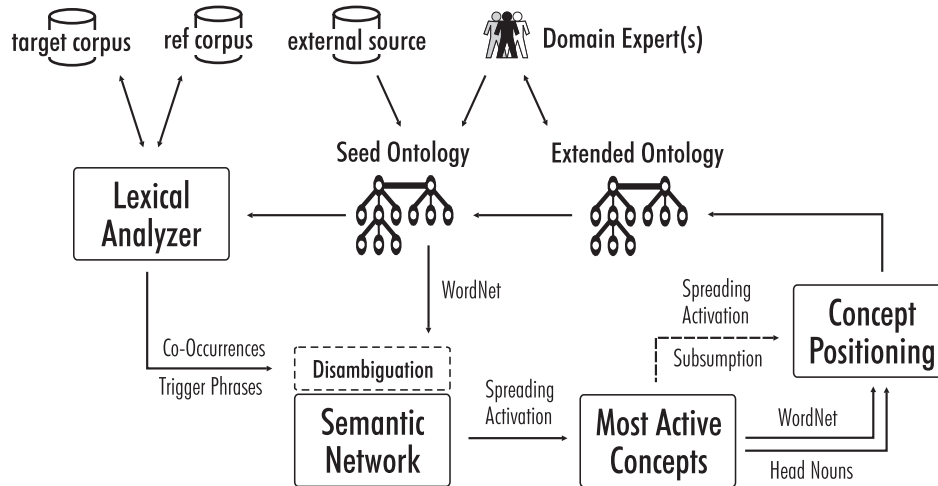


Figure 2. Ontology extension process

*OWL Lite*, a sublanguage of the *Ontology Web Language (OWL)*, provides a quick migration path for thesauri and other taxonomies [65]. IDIOM will use *OWL Lite* for a context-dependent assessment of tokens, sentences and documents. Incorporating and extending external knowledge repositories will improve IDIOM's topic detection and tracking capabilities. Potential sources include generic directories such as *TAP* [26] and *DMOZ* [57], and domain-specific thesauri such as the *EPA Terminology Reference System* [59] and the *EEA Multilingual Environmental Glossary* [58].

Given online media's dynamic character, IDIOM will continuously validate and update ontologies 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 [18]. 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* [62], *WordNet* [47; 70] and *OpenCyc* [64]. Standard ontology tools like *Protégé* [38; 67] will help incorporate suggested changes. A current pilot study aims to identify concept hierarchies using spreading activation on weighted graphs [32]. The study considers the relation of components in multi-word terms [37], 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.

Figure 2 presents the system architecture of the current prototype. A small set of terms from domain experts or from known ontology repositories is first selected as seed ontology. The seed ontology terms are then fed into the *Lexical Analyzer*. Co-occurrence analysis at both the sentence and the document level limits the influence of popular terms not related to the domain [42]. Terms are selected according to a threshold value on the co-occurrence significance. Lexical analysis is done by consulting the *WordNet* lexical dictionary [17], 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 [29].

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.

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 in 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*.

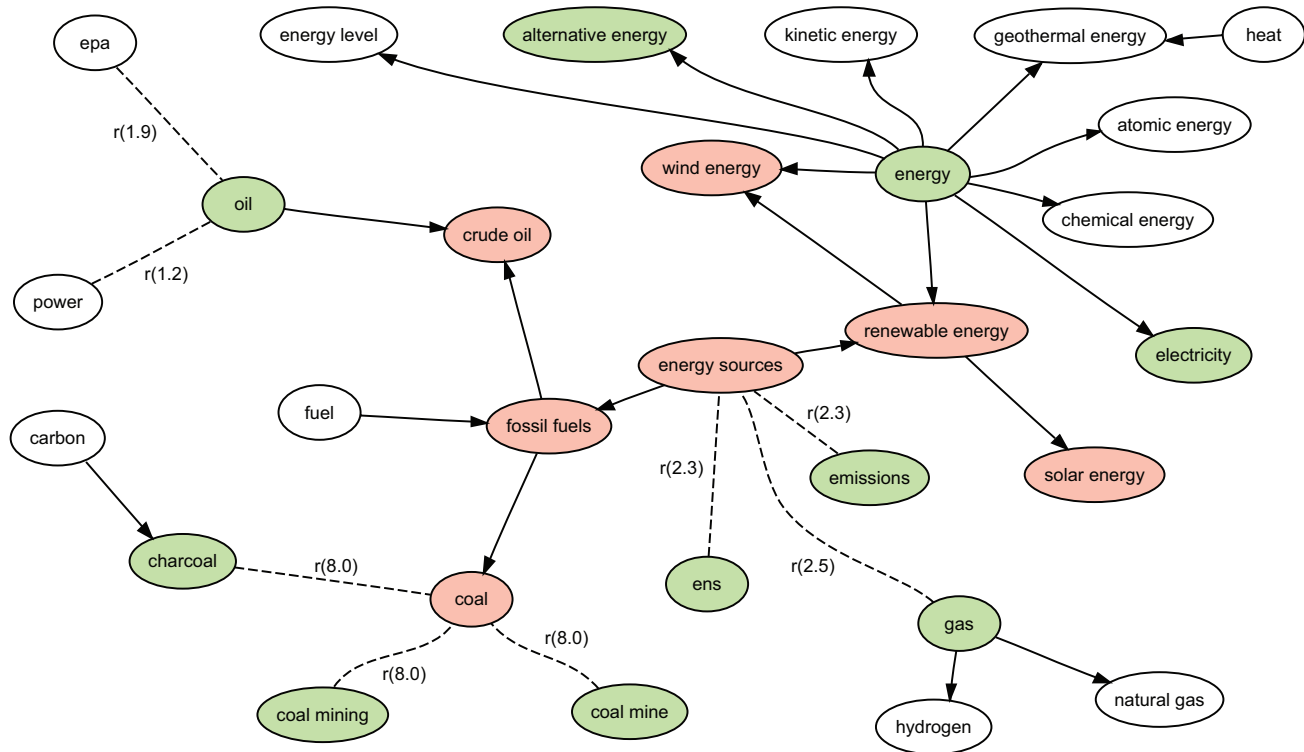


Figure 3. Concept hierarchy after two rounds of spreading activation

Hierarchically positioning the activated terms (that is, those most relevant to the domain and seed ontology), represents a challenging task. The current prototype employs three steps [32]: (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.

### 3.4. Measures of Content Redundancy

Without a proper analytical framework, content fragments often seem randomly scattered across networks. IDIOM will provide such a framework, tracking and annotating electronic documents to identify identical or similar content fragments, and to reveal fundamental mechanisms of information diffusion. Popular similarity measures use the classic *vector space model* [43], operating on vector representations that neglect ontological relationships [5]. These methods fail to detect similar mean-

ing in texts with different vocabularies. This restriction led to *latent semantic indexing* [13] and *concept space approaches* [10; 33], 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 [8; 21], a granular approach towards textual similarity popular among developers of anti-spam software [54; 56]. 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. Distributed Computing

The IDIOM roadmap presented in this paper aims to explore spatial and temporal information diffusion. While IDIOM rests on a solid foundation (see Section 3 for an overview of the webLizard technology), it will require a more 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 computing will help IDIOM

meet these demands. It will accelerate data gathering and allow sampling millions of documents by leveraging bandwidth and computational capacity of geographically dispersed systems. A distributed server cluster will 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 will pose 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 [45]: *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 [20]. *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* [7], a knowledge extraction service on top of the *Globus Toolkit* [61]. Migrating to such a distributed architecture is complex and labor-intensive. Therefore, after formally evaluating available options, IDIOM will choose 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 projects based on standard architectures.

#### 4. Conclusion and Outlook

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, the IDIOM project answers calls to identify and exploit the potential of new media for knowledge discovery and knowledge management. Our prototypical implementation of automated ontology extension [32] hints at the potential of such a comprehensive media monitoring framework.

IDIOM integrates multiple disciplines to understand the determinants, structure and impact of electronic content. A better understanding of notoriously volatile electronic content will create opportunities for organizations and individuals alike. Companies will be able to multiply the impact of their marketing and public awareness campaigns, and to accurately measure this impact. Policy makers will gain a detailed understanding of how infor-

mation replicates in interactive environments, and how this process influences public opinion.

Expected technology spin-offs include context-aware search engines, tools to build and validate domain-specific ontologies, and innovative interface technology to explore complex datasets. Unusual developments should trigger and guide the optimization of an organization's information systems, and a re-evaluation of its online strategy. 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.

Open standards for encoding and transmitting structural and textual metrics will enable other disciplines and stakeholders to leverage the gathered information. To encourage collaborative development of analytical modules, a Grid-enabled Web services layer will provide remote access to computational resources, offering raw and aggregated data in a variety of formats. Visual navigational aids based on continually refined ontologies will help manipulate and analyze these data, providing cross-references to IDIOM statistics and the underlying document repository.

**Acknowledgement.** This work is a joint project of the Research Network on Environmental Online Communication ([www.ecoresearch.net](http://www.ecoresearch.net)), Graz University of Technology, the Know-Center funded by the Austrian Competence Centers Program K+ ([www.kplus.at](http://www.kplus.at)) under the auspices of the Austrian Ministry of Transport, Innovation & Technology, Vienna University of Economics & Business Administration, and The University of Western Australia.

#### 5. References

1. Ahmad, A. and Dey, L. (2004). A Feature Selection Technique for Classificatory Analysis. *Pattern Recognition Letters*, 26: 43-56.
2. Bauer, C. and Scharl, A. (2000). "Quantitative Evaluation of Web Site Content and Structure", *Internet Research*, 10(1): 31-43.
3. Benjamins, R., Contreras, J., Corcho, O. and Gómez-Pérez, A. (2004). "Six Challenges for the Semantic Web", *AIS SIGSEMIS Bulletin*, 1(1): 24-25.
4. Berners-Lee, T., Hendler, J. and Lassili, O. (2001). "The Semantic Web", *Scientific American*, 284(5): 28-37.
5. Bernstein, A., Kaufmann, E., Bürki, C. and Klein, M. (2005). "How Similar Is It? Towards Personalized Similarity Measures in Ontologies", 7. *International Tagung Wirtschaftsinformatik*. Bamberg, Germany. 1347-1366.

6. Biber, D., Conrad, S. and Reppen, R. (1998). *Corpus Linguistics - Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
7. Cannataro, M. and Talia, D. (2003). "The Knowledge Grid", *Communications of the ACM*, 46(1): 89-93.
8. Charikar, M.S. (2002). "Similarity Estimation Techniques from Rounding Algorithms", *34th Annual ACM Symposium on Theory of Computing*. Montreal, Canada: ACM Press. 380-388.
9. Chen, H., Chau, M. and Zeng, D. (2002). "CI Spider: A Tool for Competitive Intelligence on the Web", *Decision Support Systems*, 34(1): 1-17.
10. Chen, H., Hsu, P., et al. (1994). "Automatic Concept Classification of Text from Electronic Meetings", *Communications of the ACM*, 37(10): 56-73.
11. Daft, R.L., Lengel, R.H. and Trevino, L.K. (1987). "Message Equivocality, Media Selection, and Manager Performance: Implications for Information Systems", *MIS Quarterly*, 11(3): 355-366.
12. Danielsson, P. (2004). "Automatic Extraction of Meaningful Units from Corpora", *International Journal of Corpus Linguistics*, 8(1): 109-127.
13. Deerwester, S.C., Dumais, S.T., et al. (1990). "Indexing by Latent Semantic Analysis", *Journal of the American Soc of Information Science*, 41(6): 391-407.
14. Dill, S., Eiron, N., et al. (2003). "A Case for Automated Large-Scale Semantic Annotation", *Journal of Web Semantics*, 1(1): 115-132.
15. Dill, S., Kumar, R., et al. (2002). "Self-Similarity in the Web", *ACM Transactions on Internet Technology*, 2(3): 205-223.
16. Evans, H., Atkinson, M., et al. (2003). "The Pervasiveness of Evolution in GRUMPS Software", *Software: Practice and Experience*, 33(2): 99-120.
17. Fellbaum, C. (1998). "WordNet An Electronic Lexical Database", *Computational Linguistics*, 25(2): 292-296.
18. Feng, L., Chang, E. and Dillon, T. (2002). "A Semantic Network-based Design Methodology for XML Documents", *ACM Transactions on Information Systems*, 20(4): 390-421.
19. Fensel, D., Wahlster, W., Lieberman, H. and Hendler, J., Eds. (2003). *Spinning the Semantic Web - Bringing the World Wide Web to Its Full Potential*. Cambridge: MIT Press.
20. Foster, I. and Iamnitchi, A. (2003). "On Death, Taxes, and the Convergence of Peer-to-Peer and Grid Computing", *Peer-to-Peer Systems II: Second International Workshop (Lecture Notes in Computer Science, Vol 2735)*. Eds. F. Kaashoek and I. Stoica. Heidelberg: Springer. 118-128.
21. Gionis, A., Indyk, P. and Motwani, R. (1999). "Similarity Search in High Dimensions via Hashing", *25th International Conference on Very Large Data Bases*. Edinburgh, UK: Morgan Kaufmann. 518-529.
22. Godin, S. (2001). *Unleashing the Idea Virus*. New York: Hyperion.
23. Goldenberg, J., Libai, B. and 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.
24. Granovetter, M. (1973). "The Strength of Weak Ties", *American Journal of Sociology*, 78(6): 1360-1380.
25. Gruhl, D., Guha, R., Liben-Nowell, D. and Tomkins, A. (2004). "Information Diffusion Through Blogspace", *13th International World Wide Web Conference*. New York, USA: ACM Press. 491-501.
26. Guha, R.V. and McCool, R. (2003). "TAP: A Semantic Web Platform", *Computer Networks*, 42(5): 557-577.
27. Haythornthwaite, C. (1996). "Social Network Analysis: An Approach and Technique for the Study of Information Exchange", *Library & Information Science Research*, 18(4): 323-342.
28. Hearst, M. (1997). "TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages", *Computational Linguistics*, 23(1): 33-64.
29. Joho, H., Sanderson, M. and Beaulieu, M. (2004). "A Study of User Interaction with a Concept-based Interactive Query Expansion Support Tool", *Advances in Information Retrieval, 26th European Conference on Information Retrieval*. 42-56.
30. Kiryakov, A., Popov, B., et al. (2004). "Semantic Annotation, Indexing, and Retrieval", *Web Semantics*, 2(1): 49-79.
31. Lebart, L., Salem, A. and Berry, L. (1998). *Exploring Textual Data*. Dordrecht: Kluwer.
32. Liu, W., Weichselbraun, A., Scharl, A. and Chang, E. (2005). "Semi-Automatic Ontology Extension Using Spreading Activation", *Journal of Universal Knowledge Management*, 0(1): 50-58.
33. Loh, S., Wives, L.K. and de Oliveira, J.P.M. (2000). "Concept-based Knowledge Discovery in Texts Extracted from the Web", *ACM SIGKDD Explorations Newsletter*, 2(1): 29-39.
34. Maedche, A. (2002). *Ontology Learning for the Semantic Web*. Boston: Kluwer Academic.
35. McCallum, A.K., Rosenfeld, R., Mitchell, T.M. and Ng, A.Y. (1998). "Improving Text Classification By Shrinkage in a Hierarchy of Classes", *15th International Conference on Machine Learning*. Madison, USA: Morgan Kaufmann. 359-367.

36. McEnery, T. and Wilson, A. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
37. Navigli, R. and Velardi, P. (2004). "Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites", *Computational Linguistics*, 30(2): 151-179.
38. Noy, N.F., Sintek, M., et al. (2001). "Creating Semantic Web Contents with Protégé-2000", *IEEE Intelligent Systems*, 16(2): 60-71.
39. Olsina, L. and Rossi, G. (2002). "Measuring Web Application Quality with WebQEM", *IEEE Multimedia*, 9(4): 20-29.
40. Palmer, J.W. (2002). "Web Site Usability, Design, and Performance Metrics", *Information Systems Research*, 13(2): 151-167.
41. Porter, M. (1980). "An Algorithm for Suffix Stripping", *Program*, 14(3): 130-137.
42. Roussinov, D. and Zhao, J.L. (2003). "Automatic Discovery of Similarity Relationships through Web Mining", *Decision Support Systems*, 35: 149-166.
43. Salton, G. (1989). *Automatic Text Processing*. Reading: Addison-Wesley.
44. Scharl, A. (2000). *Evolutionary Web Development*. London: Springer. <http://webdev.wu-wien.ac.at/>.
45. Scharl, A. (2004). "A Roadmap Towards Distributed Web Assessment", *Web Engineering - 4th International Conference, ICWE 2004, Munich, Germany (Lecture Notes in Computer Science, Vol 3140)*. Eds. N. Koch et al. Berlin: Springer. 171-175.
46. Scharl, A., Pollach, I. and Bauer, C. (2003). "Determining the Semantic Orientation of Web-based Corpora", *Intelligent Data Engineering and Automated Learning, 4th International Conference, IDEAL-2003, Hong Kong (Lecture Notes in Computer Science, Vol. 2690)*. Eds. J. Liu et al. Berlin: Springer. 840-849.
47. Seo, H.-C., Chung, H., et al. (2004). "Unsupervised Word Sense Disambiguation Using WordNet Relatives", *Computer Speech & Language*, 18(3): 253-273.
48. Someya, Y. (1998). English Lemma List. [http://www.lexically.net/downloads/e\\_lemma.zip](http://www.lexically.net/downloads/e_lemma.zip).
49. Song, C., Havlin, S. and Makse, H.A. (2005). "Self-Similarity of Complex Networks", *Nature*, 433(7024): 392-395.
50. Stone, P.J. (1997). "Thematic Text Analysis: New Agendas for Analyzing Text Content", *Text Analysis for the Social Sciences*. Ed. C. Roberts. Mahwah: Lawrence Erlbaum. 35-54.
51. Watts, D.J. (2003). *Six Degrees: The Science of a Connected Age*. New York: W. W. Norton.
52. Zdziarski, J.A. (2004). "Bayesian Noise Reduction: Progressive Noise Logic for Statistical Language Analysis", *MIT Spam Conference 2004*. [www.nuclearelephant.com/projects/dspam/bnr.html](http://www.nuclearelephant.com/projects/dspam/bnr.html).
53. Zeithaml, V.A., Parasuraman, A. and 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.

### Online Resources

54. *Apache SpamAssassin Project*. <http://spamassassin.apache.org/>.
55. *Business Review Weekly*. <http://www.brw.com.au/>.
56. *Distributed Checksum Clearinghouse*. <http://www.rhyolite.com/anti-spam/dcc/>.
57. *DMOZ Open Directory Project*. <http://dmoz.org/>.
58. *EEA Multilingual Environmental Glossary*. <http://glossary.eea.eu.int/>.
59. *EPA Terminology Reference System*. <http://www.epa.gov/trs/>.
60. *Fortune Magazine*. <http://www.fortune.com/>.
61. *Globus Alliance*. <http://www.globus.org/>.
62. *Merriam-Webster Online Dictionary*. <http://www.m-w.com/>.
63. *Natural Language Toolkit (NLTK)*. <http://nltk.sourceforge.net/>.
64. *OpenCyc*. <http://www.opencyc.org/>.
65. *OWL Web Ontology Language*. <http://www.w3.org/TR/owl-features/>.
66. *Plone*. <http://www.plone.org/>.
67. *Protégé Ontology Editor and Knowledge Acquisition System*. <http://protege.stanford.edu/>.
68. *W3C Platform for Internet Content Selection (PICS)*. <http://www.w3.org/PICS/>.
69. *webLyzard*. <http://www.webLyzard.com/>.
70. *WordNet Lexical Database*. <http://www.cogsci.princeton.edu/~wn/>.
71. *Zope Application Server*. <http://www.zope.org/>.