living knowledge

1st issue – April 30, 2010
http://livingknowledge-project.eu/
Feedback:
info@livingknowledge-project.eu

Livingnews

No 1

Enjoy this 1st newsletter by discovering the main topics of the LivingKnowledge project. You can also get more information by visiting our website http://livingknowledge-project.eu/ and subscribing to our news.

The LivingKnowledge project is funded by the European Commission under Project No. 231126

The overall goal of this European research project is to bring a new quality into search and knowledge management technology, which makes search results more concise, complete and contextualised.

Welcome to diversity!

The Web lives from the multitude of actors that are involved in content creation. It has achieved a democratization of content production that potentially gives a voice to everybody. However, today's search technology fails to reflect this variety in an explicit and structured way. It is the goal of the LivingKnowledge project to develop innovative methods for bias-aware, diversity-aware and evolution-aware information management and access technology to overcome this restriction.

For controversially discussed topics such as "global warming" the content available in the Web reflects the full variety of positions and their evolution over time. However, it is difficult to get a structured diversity overview, due to the sheer mass of available content and the way content is ranked by current search technology (mainly based on popularity). In addition, content is partly strongly biased without making the underlying intention explicit. A better overview over existing opinions - and support in discovering bias and analysing the underlying diversity (driven by differences in cultural backgrounds, schools of thoughts, temporal context etc.) - clearly helps in building an own opinion in a well-informed way and in reflecting and contextualizing own positions. It is the goal of the LivingKnowledge project to explore and build technology that serves this purpose. LivingKnowledge aims to make diversity a real and tangible asset of the Web.

For achieving its goal demanding research and technology development is required. Challenges have to be faced in the area of fact and opinion extraction, intelligent combination of diversity evidences found by different methods, opinion-based and evolution-aware information clustering and aggregation as well as innovative search technology which leverages bias, diversity and evolution and makes them tangible to the user. Furthermore, a thorough understanding of diversity and its impact is required as a sound foundation for the methods developed in the project. For establishing this foundation an interdisciplinary team of researchers contributes to the research and development in the FET project, which started in February 2009. In addition to establishing a sound foundation, the first year already produced exciting results in the targeted areas.

The main focus of the LivingKnowledge project is on foundational research and the establishment of a related research community. As part of this activity a testbed has been created, which fosters the experimentation of the developed methods within and beyond the consortium. Furthermore, two exemplary applications are under development, which showcase the LivingKnowledge technology. One of these applications will be coupled with Yahoo! Search technology and will explore future predictions in Web content. The second application will support the process of Media content Analysis in a professional context. ■

⬅ Partial map of the Internet based on the January 15, 2005 data found on opte.org. Each line is drawn between two nodes, representing two IP addresses. The length of the lines are indicative of the delay between those two nodes. This graph represents less than 30% of the Class C networks reachable by the data collection program in early 2005. Matt Britt (Wikipedia)

Foundations

Our goal is to study the foundations and to develop the formalisms, mechanisms and structures for effective representation and management of knowledge, which is aware of time and evolution, opinions, diversity and bias.

One of the biggest research challenges in recent years has been facing up to the emerging complexity in data, information and knowledge, in terms of size, diversity of sources, diverging viewpoints, while taking the dynamics of their unpredictable evolution in time into account. The Web is the clearest example of the enormous quantity and diversity of material – text, images and other media – continuously made available online. It is widely agreed that knowledge is strongly influenced by the diversity of context, mainly cultural, in which it is generated. Thus, while it may be appropriate to say that (some kinds of) cats and dogs are food in some parts of China, Japan, Korea, Laos and the Philippines, this is unlikely to be the case in the rest of the world. Sometimes, it is not just a matter of diversity in culture, viewpoints or opinion, but rather a function of different perspectives and goals. In fact, knowledge useful for a certain task, and in a certain environment, will often not be directly applicable to other circumstances, and will thus require adaptation. Hence the pressing need to find effective ways of dealing with such complexity, especially in terms of scalability and adaptability in data and knowledge representation.

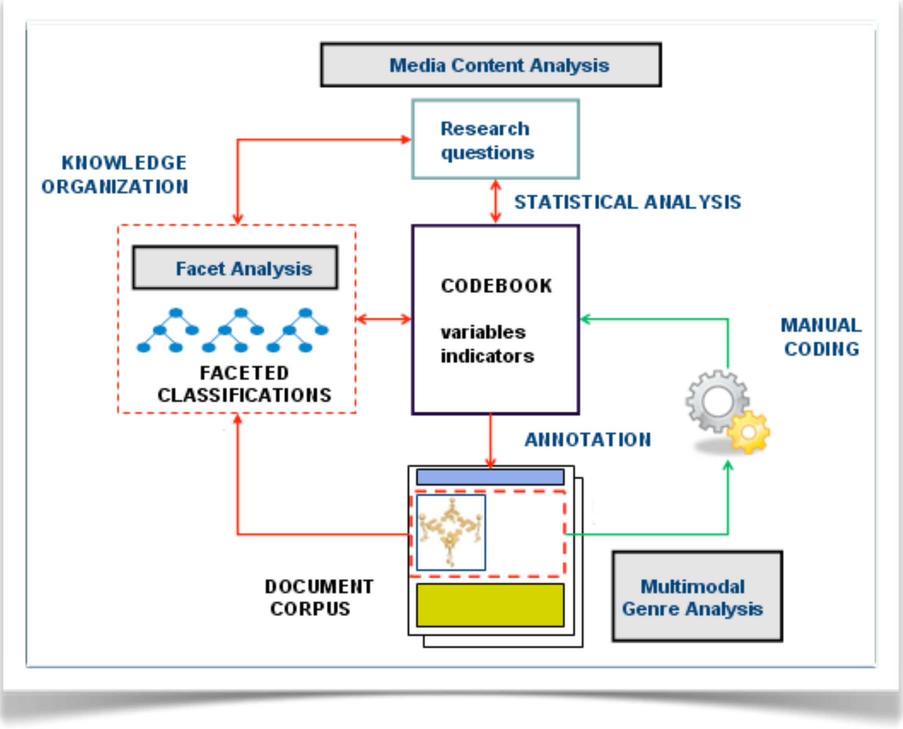
We are firmly convinced that diversity in knowledge should not be avoided, as often happens in approaches where, at design time, a global representation schema is proposed. Rather diversity in knowledge is a key feature, our goal being to develop methods and tools leading to effective design by harnessing, controlling and using the effects of emergent knowledge properties. We envisage a future where developing diversity-aware navigation and search applications will become increasingly important as they will automatically classify and organize opinions and bias producing more insightful, better organized, aggregated and easier-to-understand output by detecting and differentiating between, what we call, diversity dimensions.

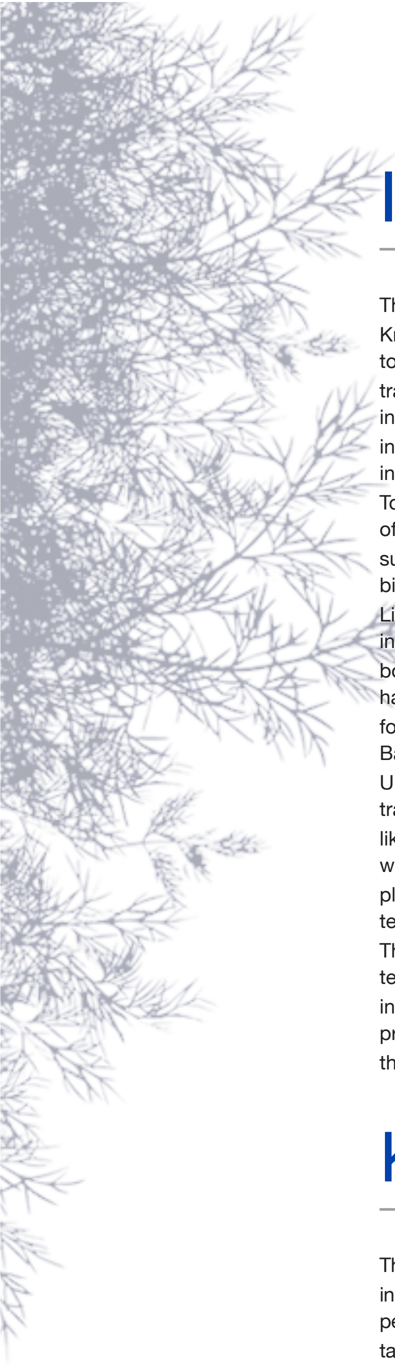
This explains our adoption of a highly interdisciplinary approach that brings together expertise from a wide range of disciplines: sociology, philosophy of science, cognitive science, library and information science, semiotics and multimodal information theory, mass media research, communication, natural language processing and multimedia data analysis. A solution to the problems posed above is gradually emerging from this synergy. In particular, we combine in a single framework the following methodologies:

- Media Content Analysis (MCA) from a social sciences perspective.
- Multimodal Genre Analysis (MGA) from a semiotic perspective.
- Facet Analysis (FA) from a knowledge representation and organization perspective.

We combine these approaches in an innovative framework in which document annotation is automated with the help of a set of feature extraction tools. For instance, the framework might parse content from different media and identify the main {concepts, people, political parties, countries, dates, resolutions, etc.} related to Immigration and which of them are the most {controversial, accepted, subjective, biased, etc.}.

The next challenges will mainly concern opinion, bias and diversity representation and management, automation of the annotation process and the implementation of the overall architecture. ■





Information extraction

The advanced and visionary functionality that the LivingKnowledge project is aiming to provide will be built upon tools that extract information from documents. This extracted information will consist of lexical and syntactic information that can be extracted from and about the text in the documents, as well as abstract descriptions of information that is hidden within the pixels of images.

Together the information extracted from these two types of media (text and images) provide an overview of the subject of the document and hint towards opinions and bias that may be within.

LivingKnowledge have technical partners who are working on the various aspects of information extraction from both text and images and in the first year the partners have developed various technologies for extracting information from documents. The University of Trento, Barcelona Media, Max-Planck-Gesellschaft (MPI) and The University of Hannover are all providing information extraction techniques from text; including basic information like whether words are nouns, verbs, adjectives, etc., as well as higher-level information such as names and places, information about events, the sentiment of sentences and the semantic structure of conceptual entities.

The technical partners in the Consorzio Nazionale Interuniversitario per le Telecomunicazioni (CNIT) are working on the detection of image manipulations which may provide clues to an intended bias within an article, and the University of Southampton and the University of Trento are both working on more general image matching and analysis techniques that may support the analysis of the text extraction partners by, for example, suggesting correlations between certain types of image use on certain types of document. The University of Southampton and MPI are also working on gathering and analysing contextual information; that is information from outside the document which provides further knowledge about entities and events explicitly given in the documents.

Individually, these technologies can only provide a small part of the overall analysis of a document and so we are now working on combining techniques and utilising the strength of many algorithms to help reinforce the accuracy of the extraction of the opinion from a document. To achieve this, a system has been built that allows the inclusion and integration of analysis algorithms and which enforces a standardised interchange format for the resulting extracted information. By the end of the first year, this system, the “testbed”, had been populated with 18 separate analysis modules and continues to grow. The testbed code then allows these analysis modules to be executed over a set of documents and a basic search engine is built on the results. As the project continues, the testbed will continue to grow both in terms of the number of document analysers available as well as the sophistication of the output. It is foreseen that the testbed will eventually provide a programming library interface on which higher level applications can be built. ■

Knowledge evolution

The constantly evolving Web reflects the evolution of society in the cyberspace. In order to understand the mutual dependencies between Web contents and the evolving societal knowledge it represents, we pursue a systematical approach toward comprehensively tracing and exploiting Web contents by: 1) natural language processing, 2) temporal fact extraction, 3) contextual image analysis, 4) classification system studies. The combination of these technologies is intended to understand the dynamics of societal events in the real world and their reflection in cyberspace.

In the area of natural language processing, automatic retrieval of opinionated pieces of text may be carried out different levels of granularity. On the coarsest level, documents are categorized as opinionated or factual (e.g. editorials vs. news). At the other end of the spectrum, methods have been proposed to carry out fine-grained subjectivity analysis on the level of linguistic expressions. We currently focus mainly on the automatic classification of individual sentences as opinionated or not. This will later pave the way for a more fine-grained analysis that can support a detailed exploration over time of the opinions held by various groups of people.

Research in the field of temporal fact extraction is based on output-oriented targeted information extraction. We gather knowledge about entities (people, companies, political parties, etc.) that evolves over time. These data are harvested in a very large semantic knowledge base called YAGO. It contains more than 2 million entities (like persons, organizations, cities, etc.) and 20 million facts about their relationships, which have been carefully harvested from Wikipedia and reconciled with the taxonomic class system of WordNet. We have recently extended YAGO toward time-awareness, which allows us to link temporal facts to any kind of Web contents that may help us to trace the evolution of opinions such as ratings, reviews, comments, news, or discussion board entries.

Work on contextual image analysis so far has concentrated on the temporal aspects of image search. This can be defined as diversity search with a temporal axis. Another area is in the provenance of images. Some news sources, particularly websites, use the same images repeatedly. Often they are to illustrate different points which may change over time. In order to exploit an image’s context we try to incorporate as much as possible contextual information captured from the camera with each photo it takes (e.g. date and/or GPS track). Once the location and time of a photograph has been determined, additional analyses can be done without the need to analyse the contents of the photo.

In the field of classification system studies our research covers two aspects. On one hand, we analyse the evolution of subject classification systems from a library classification theoretical perspective. On the other hand, we study knowledge evolution in the large from a computer science perspective, abstracting from individual entities and rather looking into the long-term changes in terminologies and topical taxonomies. We aim at the discovery of terminology shifts and their adaptation in widely used categorization schemes that reflect the collective memory/knowledge of society. ■

Bias and Diversity

The value of the Web as an information repository lies to a large extent on the diversity of the information it provides. The latter results from its open nature, further increased with the advent of Web 2.0, where more and more user-generated content is constantly made available. This diversity appears in various forms. One case is the ambiguity in keyword searches. Keyword queries often have numerous senses, and the search engine does not know which the correct interpretation in each case is. The query may also refer to a general topic, in which case the results may be related to different subtopics or aspects. It is also possible to have multiple opinions or sentiments about a topic or to have multiple types of results, such as text, images or videos. Offering diverse results to the user is critical in these situations. For ambiguous keyword queries, diversifying the results increases the probability to satisfy users with different intents. For different topics, opinions, sentiments, or perspectives, it helps the user to get a broader, overall picture of the available information, and to further explore the information space by drilling down to the results of more interest.

Typically, the results of a query are retrieved by estimating the relevance of each available document to the query, and ranking the documents in decreasing order of relevance. This means that each document is judged independently of other documents. However, this approach is not optimal when dealing with ambiguity and diversity in search, since it often results in returning documents with high similarity and overlap to each other. Then, the user becomes saturated with redundant information and eventually abandons the query. To overcome this problem, recent approaches have aimed at retrieving documents that are as dissimilar to each other as possible, in order to increase the probability that users with different intents will find at least some relevant documents in the top results.

Images also play an important role in the analysis of diversity. In some cases, images might be used along with text to try to distribute documents along a diversity axis, where the documents are primarily text based and the images play secondary roles both in the context of the document and in the analysis of their diversity. In other cases, image retrieval is the primary goal and the results of image searches need to be analyzed for diversity. As with text search, identifying and analyzing diversity in image search is important for better satisfying the user's information need when the query is poorly specified or ambiguous.

Bias detection and analysis is a new and challenging area of research. Traditionally, the analysis of media content has been performed by social scientists, searching, for example, in news archives to compare how information is reported by different sources in order to reveal bias. Methods for identifying and extracting entities, detecting topics and events, or analyzing opinions and sentiments are needed to support these tasks.

Our work in this project builds on interdisciplinary knowledge and experience to develop innovative methods and algorithms, as well as a suite of tools, to address the above challenges in detecting and exploiting diversity and bias in information. Methods for syntactic and semantic analysis of text and images are developed and applied to detect the polarity of opinions, as well as the opinion holders and their context, and various dimensions of diversity are considered, such as topics, genres, sentiments, location and time. ■

Advanced Clustering and Aggregation

Knowledge and its dissemination are strongly influenced by the diversity of cultural backgrounds, ideologies and the temporal framework. Assessments, judgments and opinions, which play an important role in many social fields such as politics and business, reflect the diversity of the different goals and points of view. A great deal of the information in the Web comes from very different sources and results in a great number of divergent views and, at times contradictions, in the facts and information displayed. In order to use and analyze the amount of information on the Web created in this way, which is increasing very rapidly, the development of a deeper understanding of diversity in the WWW as well technologies that ensure the reliable analysis of information from different sources is essential.

We address the problem of automatically structuring and summarizing heterogeneous web collections according do different dimensions of diversity including topics, temporal information, geographic context, opinions provided by the community, etc. We distinguish between three main approaches: clustering, classification, and aggregation. Classification is the process of finding a set of models (or functions) which describe and distinguish data classes or concepts, for the purpose of predicting the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known). Unlike classification, which analyzes class-labeled data objects, clustering analyzes data objects without consulting a known class label (Unsupervised Learning), i.e. class labels are not known and training data are not available. The objects are clustered or grouped based on the principle of maximizing the intracluster similarity and minimizing the intercluster similarity. That is, clusters of

■■■

objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Note that classification can be seen as a form of clustering providing additional information in form of class labels, and, on the other hand requiring training data as input. Finally, the objective of aggregation is to provide a user with a compressed representation of the above mentioned dimensions of diversity contained in a large amount of web sources; this comprises statistics, graphical representations of temporal developments, geographic representations in form of maps, etc.

We study a variety of aspects of clustering, classification, and aggregation using diverse information in different domains. Folksonomies such as YouTube or Flickr form rich sources of uploaded content from a large number of users. We exploit diverse user feedback in Web 2.0 environments to classify, rank, and annotate Web 2.0 content. Furthermore, we study the extraction of knowledge whilst taking into consideration time and geographic location as two important dimensions of diversity. We provide models for user ratings that take temporal and spatial information into account, and show results of a preliminary study in the context of movie ratings. Moreover, we built a large-scale semantic knowledge base which incorporates temporal information about semantic concepts, and can be exploited for clustering of events along a time axis. Finally, we develop approaches for providing aggregated views of blog information with the additional goal of predicting future developments. ■

Enhanced Search

We focus on developing methods for diversity-aware search and exploration of information. Diversifying search results is an important and challenging problem, required to increase user satisfaction for ambiguous queries or for queries where the results contain multiple subtopics or opinions.

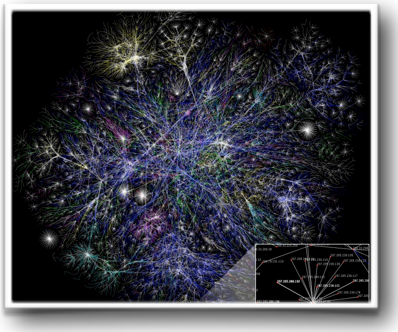
Enhanced Search refers to those technologies that aim at taking search engines beyond the traditional query-to-snippet paradigm (some times referred to as the “ten blue links”). It exploits research in information aggregation and summarization, information extraction and semantic tagging.
In our project we are specially interested in Enhanced Search possibilities to switching from document search to retrieval of factual knowledge, providing factual answers together with supporting documents; enabling searches with reference to the past, present and/or future; clustering of search results based on diversity/viewpoints; ranking of search results not only on popularity, but also on diversity and coverage of opinion.

Progresses in information extraction techniques (as outlined above) and manual annotation efforts, pursuing Semantic Web and Social Web 2.0 directions, have made it possible to begin accessing some of the entities and relations *hidden* in unstructured text. Searching based on these entities and relations promises to turn the Web into the world’s largest knowledge source. One of the earliest and more clear exponents of this approach is DBPedia (http://dbpedia.org/), which exploits the structural knowledge encoded in the Wikipedia InfoBoxes. Another example is Correlator (sandbox.yahoo.com/Correlator), which instead uses automatic information extraction techniques to analyse the raw Wikipedia text in order to provide structured faceted search interfaces over the unstructured data.

Enhanced Search is a central part of the LivingKnowledge program because of the central role that Search plays in today’s information access services. Our goal is to create innovative search, navigation, and browsing technologies dealing with bias and diversity, trust, and the temporal aspect of information.

Our work concentrates on several fronts:

- Indexing and Ranking facts and opinions: in order to search over large amounts of data it is crucial to develop the right indexing structures and the right ranking functions for these structures. Traditional search indexes are well understood today, but it is unclear how we may introduce novel more evolved forms of information such as facts and opinions, temporal profiles, authority and trustworthiness information, and more generally any form of annotation for enhanced search.
- **Query formulation techniques**: the multi-dimensionality of information will have a deep impact in query formulation as well. For example, a common query in our framework will combine a textual part with temporal constraints in natural language. Furthermore, query formulation should also support interactive and incremental refinement and the definition of constraints on other dimensions of diversity, e.g., excluding information from sources not trusted by the user; for example excluding information that comes from oil companies and related stakeholders when searching for information on global warming.



- **Navigation and browsing of information:** We will investigate into novel information visualisation techniques, which will allow the information consumer to fully profit from the technologies developed in LivingKnowledge. Based on clustering and aggregation methods, query results will be shown to the user in a summarised and easy-to-understand way and, at the same time, allow for an easy in-depth exploration of why and how the results are found and where they come from, thus enabling improved assessment and in-depth interpretation. ■