An intelligent screening agent to scan the internet for weak signals of emerging policy issues (ISA)

Joachim Klerx
AIT Austrian Institute of Technology GmbH
Foresight & Policy Development Department
Donau-City-Straße 1 | 1220 Vienna | Austria
T +43(0) 50550-4546 | F +43(0) 50550-4599
email: joachim.klerx@ait.ac.at
web: http://www.ait.ac.at

Abstract
*The political issues of the last decades have proven that there is a dramatic increase in complexity and potential damage of political decision. To better anticipate these future opportunities and threats, a variety of new methods like foresight and other forward looking activities are used in the foresight community. Being aware of these developments, the purpose of this publication is to present an **I**nternet **S**creening **A**gent (ISA) for scanning the WWW for weak signals of emerging policy issues. ISA is developed as a support software for the foresight community. With ISA the community can identify relevant topics for their weak signal assessment processes. In addition to this, ISA is developed to deliver relevance indicators to this processes and in general to supply quantitative data to assessment- , transformation- and issue interpretation processes of this community.*

The concept of ISA is, to use "wisdom of the crowds" for efficient data acquisition, issue management and community structuring. ISA is developed in a way, that uses machine learning to improve his classification rate. So the more user participate in the classification process the better ISA will get.

Keywords: Strategic planning, emerging issues, weak signals, political agent, intelligent agents, issue management, text mining

1. Introduction

The political issues of the last decades, like limited economic growth- limited resources debate, global climate change, economic crisis, terrorist threats, digital revolution and others have proven that there is a dramatic increase in complexity and potential damage of political decision. To better anticipate these future opportunities and threats, a variety of new methods like foresight and other forward looking activities are used to initiate discussions within policy on future developments. However, these activities tend to identify issues that are more or less mainstream and they do not identify emerging issues that are not yet on the policy radar. To overcome this problem, the SESTI project[1], funded by the European Commission under the 7. Framework Program, was set up. The participating partners (TNO, minOCW, JRC, AIT, MIOIR, MCST)[2] are developing a new approach for weak signals identification so that new social issues can be addressed by the policy arena in an early stage. This will be done by using different methods of early warning scanning. Whereas the SESTI project addresses a number of different concepts in early warning scanning, like wild cards, electronic weak signals, social weak signals, hypes, future trends, emerging issues [3], this paper will present a specific method for internet scanning of social weak signals, developed by AIT and used in the SESTI project as one method beside others for weak signal scan.

The overall process [3] of early warning scanning in SESTI includes a

- weak signal scanning,
- weak signal assessment,
- signal transformation
- and an issue interpretation.

This publication will describe a specific method, used in the first stage of these processes and will concentrate on WWW as data source.

The purpose of internet scanning for weak signals is to identify relevant topics for the weak signal assessment process and to deliver relevance indicators to this process, to supply quantitative data to the assessment process. The final purpose of the overall early warning scanning in SESTI is than to discover social issues that are not yet on the policy radar and can have major impact on our society. These issues will be prepared in a way that they can be discussed with high level policymakers in a systematic and productive way.

2. Emerging policy issues

The concept of "emerging issues" is often mentioned, but is not well defined from a text mining point of view. The concept is inextricably connected to semantic interpretation. In strategic

[1] http://www.sesti.info/sesti/en/project-overview

[2] TNO Institute for Applied Research
minOCW Dutch Ministry of Education, Science and Culture
JRC Institute for Prospective Technology Studies
AIT Austrian Institute for Technology
MIOIR Manchester University
MCST Maltese Council for Science and Technology

planning discussion of politicians with the research community the concept is often used to express the focus on a specific research paradigm. Obviously it is best for most researchers in this business to do research on emerging issues. So researchers tend to expect that the most important emerging issues will come up in their subject. Political lobbyists are more reliable. They know that the most important social issue will come up in their promoted subject.

In contrast to this, a topic is a more neutral term, often with a singular character. However an issue represents something that lasts longer, and needs to be taken up, to further deal with, whereas the topic does not imply, that it needs to be taken up by a politician. Generally an issue is defined as an important social question and can cause a social dispute.

As mentioned in the SESTI working paper on emerging issues [3], an established issue is already on the relevance list of frequently discussed things, whereas an emerging issue is something that is still not that widely taken up, but has a high potential of being put onto this list quite soon. This can be due to different criteria, e.g. because of its high criticality within certain contexts or because the community attracted to it gets larger. This can be done by the majority of the group or by a small power group, that has a personal interest in pushing the issue.

Thus the concept of emerging issues [3] is closely related to signals. If a signal of a priority change from a political issue remains a singular occurrence that is soon forgotten or develops into an issue is not easy to estimate at the beginning. But other observations can indicate a possible take up. If observations that relate to a specific happening that was once discovered as weak signal, aggregate, the probability rises that this cluster of observations could lead to an emerging issue. This means that after the detection of a weak signal the environment should be scanned for other happenings that relate to this. Taking this definition, an issue is not a single event, but the result of a multitude of events that tend to lead towards a certain direction.

3. Concept of weak signals

In communication theory a signal is a sign with a specific meaning to the receiver of this signal. If the communication is build up with a carrier signal of white noise, than a signal with a specific meaning has to be different from the white noise. So, as a core concept in signal processing, the signal is more or less the peak that transfers the information from the sender to the receiver. A weak signal is than a signal, which is statistically not very different to the carrier signal. Not very different can be calculated with a small but existing stochastic significance.

In our text mining process, the basic corpus, or more precise, the word frequency matrix of the basic corpus, is used as a kind of white noise (even if there is some structure) in which words have a "normal" frequency according to their meaning. In the process of weak signal discovery changes in word frequency are used to as an indicator for semantic changes.

In addition to this a second concept of weak signal will be used in scanning process. Our screening agent will use search frequency data from Google to identify an increasing social interest in specific terms. As the search frequency is very sensitive on human interest it is considered as weak signal for increasing social relevance.

4. System Architecture ISA

Building up on the concept of weak signals and the definition of emerging issues, as discussed in [3], it is obvious, that scanning the WWW for weak signals of emerging policy issues is a resource-intensive and expensive task. Google indexes about 1.3 Trillion sites, so that it is obviously not a good idea just to crawl the web and set up a text mining on this results. In addition to this there are a large number of different file formats, character sets and languages, which will cause additional problems.

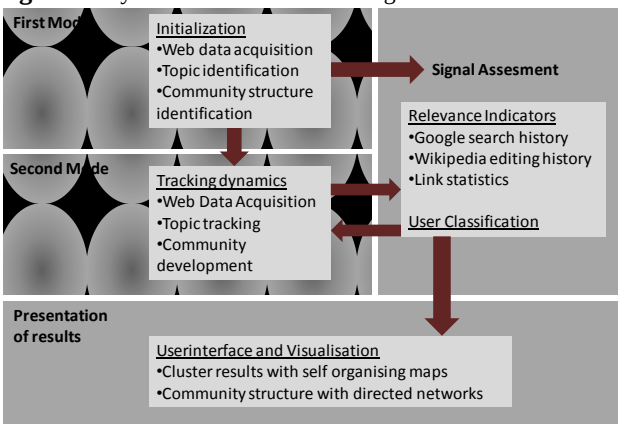
Finally, network centrality of texts in the internet is reached with a large number of inbound links, so that it is easier to find old files, with a good reputation, than new files, which is disastrous for the purpose to search emerging issues. This situation is getting worse in time.

One starting point for the engineering process was that more than 99.9% of the files on web servers are irrelevant for our issue. So the idea is to identify the relevant subset of the WWW and to use "wisdom of the crowds" for filtering and selection in the following analytical process.

James Surowiecki [14:10] mentioned "diversity of opinion (each person should have some private information, even if it's just an eccentric interpretation of the known facts), independence (people's opinions are not determined by the opinions of those around them), decentralization (people are able to specialize and draw on local knowledge), and aggregation (some mechanism exists for turning private judgments into a collective decision)" as elements for a wise crowd decision and we will show how this concept can be helpful for finding emerging issues on the internet.

The following graphic gives an overview of the overall system architecture of the **I**nternet **S**creening **A**gent ISA, which is described in the following.

Figure 1: System architecture of ISA agent



Source: AIT

As screening for weak signals of emerging issues is a dynamic task, the screening agent will work in two different modes. The first mod is developed to set up the "white noise" for the screening process and the second mod is developed to calculate the differences for each follow up screening, so that weak signals can be identified as a pattern, that is different to white

noise from mode one. In each mode, the agent will conduct a number of tasks, to build up and maintain his knowledge base.

First Mode: Initial Data Acquisition
In first mode the agent loads a specific text corpus, considered as relevant for our issue detection. In a test run, we considered each site on the internet with the phrasem "emerging issue" as relevant. A search in Google for this phrase let expect 105,000 results. Yahoo search informs their user, that they have stored about 526,003 results and on the Bing search engine, the search results statistic let the user expect 37,100 000 results (which seems strange and might be a mistake). However a normal user will only take a look at the first few results. This is very important for understanding user search strategy on the internet. A normal user take the search results as a starting point for a semantic search, which he processes like a snowball system, in which he likely will click on a link, when he expects relevant content behind the link. In a abstract sense, the agent works quite similar to the human scanning and we expect that this behavior is cost effective and efficient. Nevertheless, it is very time consuming for the user to track a larger thematic community, manually. In contrast to this, the agent can scan issue developments with high speed and thus in a short time period. A disadvantage is however, that the agent cannot understand the real semantic meaning of the sites he is crawling. So he has to rely on artificial intelligence to calculate his relevance indicators, as we will explain later in more detail.

First, the agent makes use of wisdom of the crowds in a way, that he uses the yahoo search engines relevance algorithm to download a list with highly relevant links for our issue. As our potential text corpus on the internet contains hyperlinks, the text corpus can be thought of as a directed network, with authorities and hubs, whereas an authority node is a site with a lot inbound links and a hub is a site with a lot of out bound links. Due to the relevance algorithm of the search engine, the authorities are higher ranked than the hubs, so that we get basically a list of authorities, related to our specified search strategy.

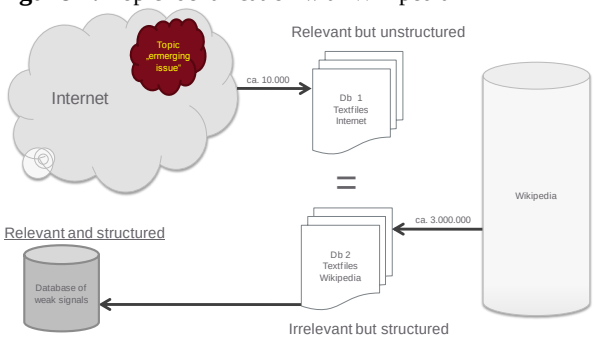
Next, the agent follows each link, which was extracted from the search engine result list texts and downloads the corresponding text information. In case, that this text contains the search string, ISA, extracts title, keywords, main text and links from this site and writes the results to his database. This step is a mixture of human search strategy and automatic crawling. In traditional automatic crawling you download all the sites which are reached by a specific crawling rule, e.g. all links from a specific domain or all links up to a certain threshold. In human crawling usually only the links are used, which let the user expect a highly relevant result, according to some kind of semantic interpretation of the user. Up to now our agent follows the links, where the text of a site contains the search string. In future version of ISA we plan to increase the “intelligence” of the agent by using text classification algorithm like Naïve Bayes or Support Vector Machines, so that the agent can learn from human user behavior. For the first version it was enough to keep this as simple as possible.

In a third and final step of data acquisition, the agent follows all extracted links, extract the site attributes and test whether the main text of the site contains the search string. To prevent the agent from “black holes” for internet crawler, the agent will not download more than about 100 documents from a single domain. All text results are grouped by domain, so that there is a consistent domain –text/date relation in the database. This database forms our data source for topic identification as next step and network analysis as follow up analytical.

First Mode: Initial Topic Identification
For topic identification, we use the text corpus from initial data acquisition, enriched by organization names from whois databases, so that each text information is related to a specific domain and thus to a specific organization (the owner of the domain). From this data set topics are identified with a cluster analysis from the word vectors of each text.

The problem is that it is not possible to generate topic title and topic description for each cluster automatically. Thus generally speaking, the dataset is expected to be relevant but unstructured in a sense, that there is no title and no description for each topic. Wikipedia articles in the opposite are structured with title and description but most of them are obviously irrelevant for our purpose. So the idea is to combine a topic text from cluster analysis with title and description from Wikipedia. This results in a structured list of well-known issues for the addressed community. The whole process is symbolized by the following graphic.

Figure 2: Topic identification with Wikipedia DB



Source: AIT

To combine our cluster with Wikipedia articles it is necessary to calculate a distance measure for each cluster-article relation. In May 2010 we found around 3.5 million articles in the English Wikipedia database. So for each cluster it is necessary to calculate 3.5 million distance measures from their word frequency matrixes, to find out which article fits best to the cluster, which limits the number of cluster. After this, the agent has a cluster – title - description relation in his database. With the relation text id – organization from the earlier mentioned whois request it is possible to map the identified cluster topics with organization.

This dataset can be interpreted as “white noise” of well known issues in a specific community, on which new signals might come up, e.g. when the topics attract more organization or disappear over time. However, to find weak signals for emerging issues, some analytical steps are necessary. The concept of “weak” implies that there are small but important changes in time with high impact. So, the following steps of community structuring will help to focus on the relevant signals.

First Mode: Initial Community Structuring
A network analysis is set up with the existing data set from the text mining process to enrich the results of the topic identification. For detection of weak signals it is helpful to connect topic dynamics with organization, their community position and their interests. Therefore we use hyperlink

information from our database to calculate a directed network of organization, with domains as nodes and the amount of hyperlink as connection between the nodes.

Each hyperlink from domain A to domain B is interpreted as vote for trust from domain A to domain B. As a result we get a kind of community network in which the domains, which are related to (owned by) a specific organization, can have a specific position (hub and authority position). As each node is although a member of a topic cluster, it is possible to visualize each thematic cluster in the network with a specific color. We expect that position changes within a thematic subnet or across, will give hints to a shift in community attention to specific topics, which can be interpreted as weak signals for upcoming issues. However, we are aware of the fact, that changes are only visible, if the whole process of data acquisition, text mining and network analysis is at least performed twice. So we do not have results for network dynamics in the first test run.

However up to now we saw, that the network structure is quite different for our test search strategies. While in the text corpus from the search strategy “emerging issue” e.g., the community network has quite a lot of inbound links between community members, the text corpus from the search strategy “human enhancement” has more outbound links to homepages with no “human enhancement” on the site. Up to now, we only recognize these differences. Further empirical experience will show, whether these patterns can be used to identify emerging issues.

Second Mode: Data acquisition, Topic tracking and Community development

The development of the second mode is not finished yet so the description is based on the theoretical concept and is up to changes due to practical experience. Basically the second mode is a variation of the first mode with data acquisition, topic identification and community identification and their corresponding dynamic variations topic tracking and community development.

Data acquisition is more or less equal to initial data acquisition. However to discover changes, it is necessary to download the whole data set again for each period in time, as there is no reliable version flag standard in html, up to now. Thus the agent will proceed with the whole mode 1 download for several times (search in a search engine, download full text from search results and follow all extracted links so that all links with search strategy string in text are downloaded). After each download the corpus is treated in the same manner than in mode 1, with word vector creation and clustering.

Topic tracking is different to topic discovery in a way that it is necessary to now the topics that should be tracked and that for tracking additional topic attributes are used. For tracking basically four events of topic-development are of interest. Topics might get more important or less important and topics might arise or disappear. For the first two events it is necessary to define an "importance" - indicator, e.g. the number of domains or organization in the topic or the external links to topic sites. For the second two events it is necessary to calculate difference measures, based on the reference word vector from the Wikipedia description of the topic and based . If the distance measure exceeds a specific threshold it is assumed, that a new topic did come up and the topic identification algorithm will be executed.

Community development is based on the calculation for initial community structuring with our network model, with organization as nodes and directed links as edges between the nodes. In this network, some organizations have privileged positions, like hubs or authorities, as mentioned in the description of the initial community structuring. These positions might change over time and one organization might become more important than other. These changes will be analyzed in the community development.

All these dynamic measure concepts are set up to find weak signals for emerging issues. However we expect to see first the strong signals for emerging issues and only with improvements in the sensitivity of the measure concept, we expect to find more and more weak signals. In addition to these internal indicators an evaluation module with external indicators is part of the internet screening agent ISA.

Evaluation module: Signal Assessment

In addition to the internal evaluation methods of the second Mode, the evaluation module is developed to offer external indicators for evaluation. The most important question for the agent is: "Which pattern makes a signal to a weak signal for an emerging issue?". The purpose behind the concept of weakness is to detect changes as early as possible so that the detection of weak signals is not practiced as an end in itself. For signal assessment it is more important to draw the attention to the time frame and to concentrate on the detection early in time than to draw the attention on weakness. Basically emerging issues in a community can be discovered by looking on the communication behavior in the community or by looking on the knowledge search behavior in the community. We are using both strategies with the following external indicators. The agent will reference them as additional attribute for each identified topic

The agent will use editing statistics from Wikipedia as the first external relevance indicator. The idea draws upon the well known fact that some of the articles in Wikipedia are changed very often and others are not. This is often due to conflicting interests in a specific topic or a wider interest in the topic, which we think, can both be interpret as an indicator for social relevance. If an upcoming issue let expect a high impact on society it might be a weak signal for an emerging issue.

As second external relevance indicator, we use the usage pattern from Google to identify topics with increasing or declining attention. As the search statistics reflects the world wide search behavior, it can be expected that an increase in search term usage reflects an increase in overall social relevance of the topic that is referenced by the search strategy. So we use our topic title from Wikipedia as search strategy in Google Trends and the agent loads the search statistics from these topic titles. With this relevance indicator, the agent finds out, when a topic gets attention of a worldwide community. Again this is stored as an attribute for each identified topic.

As an outlook for future development direction for signal assessment, we are developing the agent in a way that user classification can be used to improve automatic signal classification in the future. With this, we expected that the agent will develop the capability to find unknown pattern for emerging issues. However we are aware of the fact, that this signal assessment with mechanical learning algorithm needs text samples that are tagged by the user, according to their

relevance. Thus this feature will only work if there is an existing user community.

5. Experiences from the first test run of the ISA software

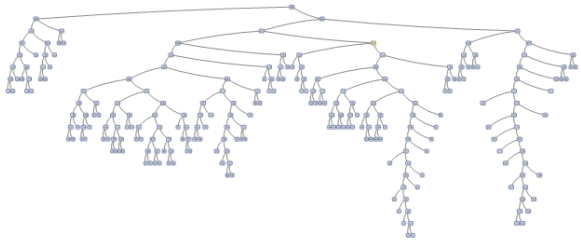
The first test run of the internet screening agent ISA has shown that the engineering concept in general is operational. However some weaknesses have come up, in most of the agent modules, which will lead to an improvement in the next version of ISA. In the following we will present some of our insight results, which are produced by the agent to build up his knowledge base.

Crawling and spider are working as expected. The results make clear, that in addition to the html format, the pdf format is very important at least for scientific communities. So it is on our agenda to include pdf parsing in the data acquisition module. We are aware of, that parsing pdf files will have an effect on network calculation results, as pdf files usually do not include hyperlinks. So we expect that it is necessary to mark them as pdf-source and exclude them from network calculation.

A small but very important detail in the spider process has come up as an important part for the scalability of the agent. The html download and parsing process is scalable because it is easy to set up a thread for each download and parsing process. However to prevent the spider from importing sites more than once, it is necessary for the spider to check in the database, whether the actual url in queue exists already in the agents database or not. This is a task that will need an increasing amount of time, when the list of urls in queue becomes large. The execution time for this single database request will grow potentially for quite a while for a typical crawling strategy, as each parsing of a new site will add more than one link to the spider queue. So up to now it is an open problem to increase the speed for this database request.

In the topic identification module the results are even better than expected. For practical reason we use two different clustering algorithm. For initial topic identification, we use an unsupervised hierarchical clustering with mixed Euclidean distance measure. The following figure 3 shows a typical result for a small corpus of our spider results for the search strategy "emerging issue".

Figure 3: Unsupervised hierarchical clustering of the spider results "emerging issue"



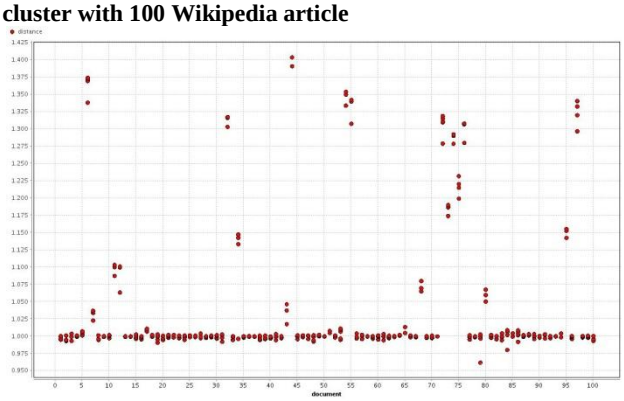
Source: AIT

The results show that the cluster algorithm is capable to identify different topics and the corpus might contain 4 different topics. In a next cluster process we use this information to identify and label these topics and to calculate the word vectors for each topic, so that we have a cluster id –word vector information in the database.

To identify and tag each cluster with a human readable title and a topic description we use an k-means distance algorithm to identify the Wikipedia article, that fits best to each cluster.

Figure 4 shows the inverse distance (larger is more similar) for one cluster with 100 Wikipedia article (for visualization purpose, we limited the results to 100).

Figure 4: K-means similarity plot of one "emerging issue" cluster with 100 Wikipedia article



Source: AIT

It is quite clear, that most of the Wikipedia articles are not similar to our cluster example (distance nearby 1). Some of the articles are more similar than the average and only a few are quite similar. So the topic identification algorithm is working as expected. The agent is looking for the article with maximized inverse distance and he stores this article as descriptor for the corresponding cluster topic and thus has a topic with title and description in his database. With a whois web service it is possible to identify all the organization with a specific domain, and with the agent database we know which domain is engaged in a specific topic. From this it is possible to start the community calculations.

The purpose of the community structuring module was to get indicators about the position of all organization in a specific topic. Up to now, the network calculation and visualization in the module for community structuring is an open task. So there are no results for now, but we do not expect too much problems in this module, as it uses well known methods.

To sum up our results from the first ISA scanning process, we got the impression, that it is not an impossible task to develop software agents that can help in scanning for emerging issues. However the key problem in scanning for emerging issues is, that computer have no semantic understanding of what is important, relevant, interesting and so on, so the computer has to rely on statistical indicators and it is inherent difficult to discover weak signals for new or emerging issues as typical statistical pattern like maximum, minimum, average and so on are not typical for weak signals. Nevertheless human scanning is very resource intensive and comes to a natural limit with very small source documents. So it seems that both, automatic agent scanning and human scanning work quite different and both have their advantages and disadvantages.

6. Conclusion and future perspectives

As already mentioned, the engineering concept seems to fulfill the expectation and we expect, that the concept for a weak signal scan has some advantages in comparison to the more often used manual scan. We already discussed the advantages and disadvantages of automatic internet scan with our intelligent agent. Even after we tried to adapt our agent to human search strategies there remained a number of fundamental differences.

Human scanning for weak signals for emerging issues is based on a specific understanding of political issues and thus is based on semantic understanding of the issue and human intuition. Usually humans tend to search like a snow ball system, from one to the next document in the WWW. They might skip a full new cluster if the content of the first few documents of the cluster are not appropriate.

In this article was shown, that automatic scanning is possible, but is based on statistical methods of machine learning and text mining. This is by far not an equal replacement for the human scanning but it might be a supporting method to overcome the drawbacks from human scanning.

ISA was developed as scanner for policy issue and due to the specific sources it is optimized for this subject. However in the future, we expect that ISA can be helpful for scientific issue scanning although. Especially community identification and structuring is applicable to epistemic communities where the same question arise, like who is referenced by whom, what are the most important subjects and which subjects are getting more important in this community. The next more user friendly version will show whether ISA will be accepted in the foresight community or in other scientific communities.

7. References

[1] Albeverio, S., Jentsch, V., Kantz, H., **Extreme Events in Nature and Society**, Universität Bonn, Interdisziplinäres Zentrum für Komplexe Systeme und Institut für Angewandte Mathematik, 53115 Bonn, Germany, Springer Verlag, 2005

[2] Bun, Khoo Khyou and Ishizuka, Mitsuru, **Emerging topic tracking system in WWW**, Knowledge-Based Systems 19 (2006) 164-171, Elsvier, 2006

[3] Butter, Maurits, et al., **Early warning scanning; dealing with the unexpected, An operational framework for the identification and assessment of unexpected future developments**, SESTI working paper, 13. December 2009

[4] Dillard, Logan, **Valence-Shifted Sentences in Sentiment Classification**, University of Washington, Institute of Technology - Tacoma, 2007

[5] Esuliand, Andrea and Sebastiani, Fabrizio, **Page Ranking WordNet Synsets: An Application to Opinion Mining**, Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pages424–431, Prague, Czech Republic, June 2007

[6] Feinerer, Ingo, **A Text Mining Framework in R and Its Applications**, Wirtschaftsuniversität Wien, Department of Statistics and Mathematics, Vienna, Austria, 2008

[7] Hiltunen, Elina, **Good Sources of Weak Signals: A Global Study of Where Futurists Look For Weak Signals**, in Journal of Futures Studies, May 2008, 12(4): 21 – 44, Finland Futures Research Centre, Finland, 2008

[8] Lee, Changki, Lee, Gary Geunbae and Jang, Myunggil, **Dependency structure language model for topic detection and tracking**, Information Processing and Management 43 (2007) 1249–1259, Elsevier, 2007

[9] Lee, Changki, Lee, Gary Geunbae and Jang, Myunggil, **Use of place information for improved event tracking**, Information Processing and Management 43 (2007) 365–378, Elsevier, 2007

[10] Manu Aery, Naveen Ramamurthy, Y. Alp Aslandogan, **Topic Identification of Textual Data**, Technical Report CSE-2003-25, Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX 76019, 2003

[11] Narayanan, Arvind and Shmatikov, Vitaly, **Robust De-anonymization of Large Sparse Datasets**, University of Texas at Austin, 2008

[12] Roth, Camille and Bourgine, Paul, **Epistemic communities: description and hierarchic categorization**, Center for Research in Applied Epistemology, CNRS/EcolePolytechnique, 1 rue Descartes, 75005 Paris, France, Dec. 2004

[13] Russell, Stuart & Norvig, Peter, **Artificial Intelligence, A Modern Approach**, Prentice Hall Series in Artificial Intelligence, Pearson Education, 2003, USA

[14] Surowiecki, James, **The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations**, Anchor, New York, 2004

[15] Taleb, Nassim, **The black swan: the impact of the highly improbable**, Random House, New York, USA, 2007

[16] Wooldridge, Michael & Jennings, Nicholas R., **Pitfalls of Agent-Oriented Development**, Department of Electronic Engineering, Queen Mary & Westfield College, University of London, London E1 4NS, United Kingdom, fM.J.Wooldridge@qmw.ac.uk; N.R.Jenningsg@qmw.ac.uk, 1998

[17] Zdravko Markov and Daniel T.Larose, **Datamining the WEB, Uncovering Patterns in Web Content,Structure and Usage**, Central Connecticut State University, New Britain, CT, Wiley-Interscience, 2007