# City Research Online

## City St George's, University of London

**Citation:** Danilova, N. & Stupples, D. (2012). Application of Natural Language Processing and Evidential Analysis to Web-Based Intelligence Information Acquisition. Intelligence and Security Informatics Conference (EISIC), Proceedings, 2, pp. 268-273. doi: 10.1109/eisic.2012.41

This is the unspecified version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

**Permanent repository link:** https://openaccess.city.ac.uk/id/eprint/1603/

**Link to published version:** https://doi.org/10.1109/eisic.2012.41

**Copyright and Reuse:** Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to City Research Online policy.

City Research Online: http://openaccess.city.ac.uk/ publications@citystgeorges.ac.uk

# Application of Natural Language Processing and Evidential Analysis to Web-Based Intelligence Information Acquisition

Natalia Danilova, David Stupples
School of Engineering and Mathematical Sciences
City University London
London, United Kingdom
{natalia.danilova.1, d.w.stupples}@city.ac.uk

*Abstract* — **The quality of decisions made in business and government relates directly to the quality of the information used to formulate the decision. This information may be retrieved from an organization's knowledge base (Intranet) or from the World Wide Web. Intelligence services Intranet held information can be efficiently manipulated by technologies based upon either semantics such as ontologies, or statistics such as meaning-based computing. These technologies require complex processing of large amount of textual information. However, they cannot currently be effectively applied to Web-based search due to various obstacles, such as lack of semantic tagging. A new approach proposed in this paper supports Web-based search for intelligence information utilizing evidence-based natural language processing (NLP). This approach combines traditional NLP methods for filtering of Web-search results, Grounded Theory to test the completeness of the evidence, and Evidential Analysis to test the quality of gathered information. The enriched information derived from the Web-search will be transferred to the intelligence services knowledge base for handling by an effective Intranet search system thus increasing substantially the information for intelligence analysis. The paper will show that the quality of retrieved information is significantly enhanced by the discovery of previously unknown facts derived from known facts.**

***Keywords – information intelligence; natural language processing; semantic similarity; evidential analysis; grounded theory***

## I. OVERVIEW TO THE APPROACH

The quality of decisions made in business and government correlates directly to the quality of the information used to formulate the decision. Most of the information used for intelligence analysis will, in the future, be harvested from the Web as this is becoming the richest source. An Intelligence service Intranet held information (its knowledge base) can be efficiently manipulated by enterprise search systems based upon either semantics such as ontologies, or meaning-based computing. These technologies imply comprehensive (and often automatic) indexing and tagging of the Intranet knowledge base textual information. Existing Web, as originally described by Tim Berners-Lee in 2001 [1], was expected to evaluate into Semantic Web, that encourages simply the inclusion of semantic content in Web pages, making it not only human readable, but also machine readable. However, most of the current Web remains poorly semantically tagged, making it impossible to apply effective enterprise search methods to Web-based intelligence information extraction. If the Web is to be used for improving decision-making, then new more effective search methods must be developed in order to collect and correlate the best information. This new search method may be used to harvest Web data in accordance with carefully controlled parameters and transferred to the Intranet knowledge base where upon enterprise search technologies may be then applied in the usual way.

It should also be noted that an Intranet knowledge base can become too historic and Web-based knowledge more effectively reflects the current state of the world. Regular updates to an Intranet knowledge base would make sense.

Donald Rumsfeld [2] stated (paraphrased): "there are 'known knowns' (KKs) – that is things we know we know; there are 'known unknowns' (KUs) – that is some things we know we do not know; but there are also 'unknown unknowns' (UUs) – that is things we don't know we don't know." Effective decision-making requires trusted, focused and relevant information. We should be comfortable with both 'KKs' and 'KUs', as these are straightforward to find. The problem being that much of the rich information required for good decisions may be in the category of 'UUs'. So the important question to be asked is how we find the relevant 'UUs' to enrich and improve decision-making? In effect we need to identify an enterprise search solution equivalent for the Web that can handle the vast amounts of information involved and in the very many different format types. This equivalent, what may be categorized under the collective title of evidence-based NLP, is the subject of this paper.

Evidence-based NLP may be considered as comprising three integrated processes that are as a whole iterative. Firstly, the application of NLP methods to enable the filtering of Web-search results to form a set of relevant information, thus overcoming the search engine keyword and ranking mechanisms that limit the use of a search engine approach.

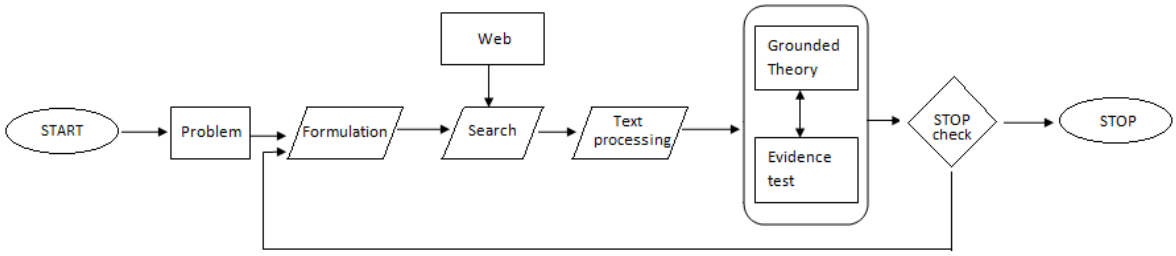


Figure 1. General level process flow diagram

Thus, the captured sets of 'KKs', 'KUs' and 'UUs' are semantically related and, therefore, relevant to the topic being considered.

Secondly, this captured set is subjected to the application of Grounded Theory where 'UUs' are specifically identified and used to test the completeness of the evidence.

Thirdly, the application of the Evidential Analysis is used to test the quality of gathered information and hence setting a quality parameter for the efficacy of the eventual decision-making process.

The three processes together are applied iteratively to the Web with an expanding query base using converted 'UUs' in order to identify the best information for the target decision process. Development of three processes together with a specifically design evaluating case study will form the structure of the paper. A discussion of the results will be used as a conclusion.

## II. Applying Natural Language Processing

The traditional approach to the Web-search is based on indexing of the Web content, building an index database, and then searching for the keywords that match the content of this database. However, this strategy will not easily support intelligence information acquisition. The Google search engine (the most commonly used) is able to find several millions of Web-pages and display up to 1000 results for a particular search in a fraction of a second, but these pages are not necessarily semantically related. Even though Google currently has the best duplicate content filtering technology [3, 4], it cannot analyse the meaning of the texts to eliminate semantically repeated documents, quantity does not always mean quality.

Fig. 2 illustrates the dependence between the size of information pool and its quality. The quality of decisions depends on the quality of information. The aim for the intelligence service analysts will be to collect as much relevant information as possible, thus not exceeding the optimal amount of information that causes information overload and hence reduces the quality of the decision as a result.

The keyword matching technique essentially misses important information, while ranking strategy may place irrelevant search results at the top of the list. What should also be borne in mind is that the keyword being used reflects what the author has in mind and not necessarily what is required by the intelligence search, resulting in possible relevant information being missed.

A recently suggested approach to overcome this information problem is 'concept search', i.e. analysis of unstructured (plain) text for information that is conceptually similar to the information provided in a search query; ideas expressed in the retrieved information are relevant to the ideas in the text of the search query. Concept search is widely used in enterprise-search and data management systems, such as Autonomy [5], that operate with the finite knowledge base, making it possible to "understand" the meaning of the short query by extracting the meaning of the documents that are currently opened on an analyst's PC desktop. Regardless of the effectiveness of such methods in the Intranet environment, Web scale far exceeds the amount of information that these methods can process reasonably in a realistic timeframe.

The new approach, proposed in this paper, supports Web-based search for intelligence information acquisition. The proposed solution follows the steps shown in the diagram above (see Fig. 1). Text processing unit extracts the Web-pages that are relevant to the initial knowledge base content. Grounded theory is used to test the completeness of the knowledge base, while evidential analysis test the quality of gathered information. Once the quality and completeness processes have approved the search content, the data files containing the correlated Web-search information can be transferred to the intelligence service knowledge base for further analysis.

To explain, initial target knowledge and search objectives are identified manually by intelligence or business analysts and presented in an unstructured text format. This target knowledge directly relates to the collection of facts and information to enable a more formal definition of the topic. This collection forms the initial set of 'KKs' and is considered as base evidence or initial knowledge on the topic. The larger set of 'KKs' at this stage may ensure a better result although the quality of 'KKs' is important.

It is quite likely that the queries for the Web-search will be formulated by analysts working within the intelligence community. This ensures that the Web-retrieval will augment the intelligence service knowledge base, hence maintaining integrity and consistency, and update accordingly.

The text in the initial knowledge base is processed in order to filter out stop-words – the most commonly used English words, superfluous with respect to our needs. Search objectives relate to 'KUs' and, thus, form the initial queries for the search engine. The initial 'KUs' are identified by analysts probably in workshop sessions and are generated from existing intelligence gap analysis.

Use is then made of a traditional search engine, such as Google, since it employs the largest index base. The aim is to build not only accurate, but also complete evidence; the search engine should not skip a Web source because it is not in its index base. It is more prudent to filter unrelated text at a later stage.

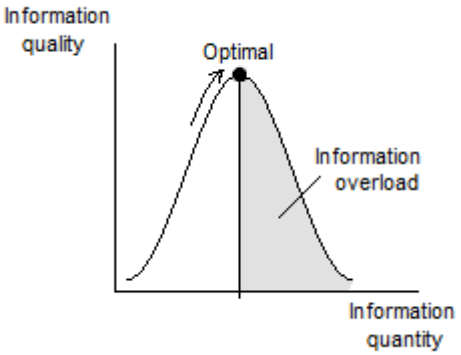


Figure 2.Information quality vs. information quantity

Clearly, the Web contains a vast amount of valuable information. However, in practice, due to the complicated and flexible layout, the main content of a Web-page is usually surrounded by noisy information (such as menu, header, advertisement, etc.). Therefore, extracting the main text of a Web-page is a critical processing task, if relevant intelligence information is to be identified. Hitherto, there have been a number of researches conducted on eliminating noisy information from Web-pages [6, 7, 8, 9, 10]. For this experiment NLTK 2.0 package for Python (http://nltk.org) is used to eliminate noise and extract header sections of pages.

Once the text has been extracted, it also needs parsing to eliminate stop-words. There are several stop-word cancelling techniques [11, 12, 13] traditionally used in NLP applications. Although, usually a stop-word list is domain depended, for the experiments we used a classic list of 250 stop-words in English suggested by Van Rijsbergen [11] that is often used as a test baseline.

Our initial knowledge text (base evidence) and the collection of texts from Web-search results are now presented for semantic analysis. The aim of this stage is to filter out those Web-pages that are semantically related to the initial evidence of current search iteration. This research firstly uses a hybrid approach developed by Hirst & Mohammad [14] that combines the co-occurrence statistics with the information in a lexical source, and employs a distributional measure of concept-distance by calculating the distance between the distributional profiles of concepts rather than words. Concepts in this case refer to the meanings of words; different words can belong to the same concept. For example, the words COFFEE and TEA belong to the concept BEVERAGE. The distributional profile of a concept is the strength of association between it and each of the words in its context. The context of a word was considered as all the words that are within the text window of ±5 words, i.e. 5 words to the left from the target word and 5 words to the right. The closer the distributional profiles of two concepts, the smaller is their semantic distance. For the lexical source we use Roget's Thesaurus (www.roget.org) that, in contrast to traditionally used WordNet [15], classifies all English words into 1044 categories.

Based on a detailed survey of semantic distance measures (see [14]), we have chosen the adapted Cosine method to estimate distributional distance between two concepts. The choice of the Cosine concept distance measurement was made based on the highest level of correlation with human rated word pairs of automatic rankings [16]. The Cosine distributional distance measure is denoted by:

$$Cos(c_1,c_2)=\frac{\sum_{w\in C(c_1)\cup C(c_2)}(P(w|c_1)\times P(w|c_2))}{\sqrt{\sum_{w\in C(c_1)}(P(w|c_1))^2}\times\sqrt{\sum_{w\in C(c2)}(P(w|c_2))^2}}, \quad (1)$$

where $w \in C(c_1) \cup C(c_2)$ is the set of words that co-occur with concepts $c_1$ and $c_2$ within a text window of ±5 words in both texts. Thus, (1) measures the semantic distance between each concept in each text, and treats the distributional profiles of concepts as vectors of the size equal to the number of all unique words in both texts. $P(w|c_1)$ and $P(w|c_2)$ are the conditional probabilities of a word $w$ co-occurring with any word listed under the category $c$ in the thesaurus. Conditional probabilities are used as strengths of association between each word and each concept in both texts, and are taken from the distributional profiles of concepts. The value for Cosine measure in our case lies between 0 and 1, indicating semantic remoteness of two concepts when the value approaches 0 and semantic closeness when the value is close to 1.

The use of thesaurus categories as concepts allows pre-computing of all concept distance values required in a form of concept-concept distance matrix of a size much smaller than word-word distance matrix.

Having the concept distances, we then calculate similarity of evidence text and texts from Web-search results list. We have adapted the formula for measuring similarity between texts, proposed by Corley et al. [17]. Their original method measures the semantic similarity of texts by exploiting the information that can be drawn from the similarity of the component words. This research adapts their method by involving concept-to-concept distance instead of word-to-word distance to measure semantic similarity between two texts. Given a measure for semantic distance between each of the concepts in each of the texts, it is possible to define the semantic similarity of two texts – the initial knowledge base text $T_1$ and the candidate text $T_2$ using a metric that combines the semantic similarities of each text in turn with respect to the other text. The similarity between the two texts $T_1$ and $T_2$ is therefore determined using the following function:

$$sim(T_1,T_2)=\frac{1}{2}\left(\frac{\sum_{c\in\{T_1\}}(\max Sim(c,T_2)*idf(c))}{\sum_{c\in\{T_1\}}idf(c)}+\frac{\sum_{c\in\{T_2\}}(\max Sim(c,T_1)*idf(c))}{\sum_{c\in\{T_2\}}idf(c)}\right). \quad (2)$$

First, for each concept $c$ in the initial knowledge base text $T_1$ we identify the concept in the candidate text $T_2$ that has the highest semantic similarity ($\max Sim(c,T_2)$), according to the concept-to-concept similarity (Cosine measure) described above. Next, the same process is applied to determine the most semantically close concepts in $T_2$ compared to the concepts in $T_1$ $\max Sim(c,T_1)$. The concept similarities are then weighted with the corresponding concept inverted document frequency $idf(c)$, that has the value 1, if the concept $c$ is used in both texts, and 0.5, if the concept is used only in one of the two texts. Next, the concept similarities are summed up, and resulting similarity scores are combined using a simple average.

This text similarity score has a value between 0 and 1, with a score of 1 indicating identical texts, and a score of 0 indicating no semantic overlap between the two texts. Once the text similarity score is defined, we can then filter the gathered knowledge (evidence) by leaving those pieces of text

(corresponding to Web-search results) that show the similarity value below the chosen threshold of 0.5. We can then add the remaining texts to the existing evidence base.

## III. Applying Grounded Theory

The information search process is independent of the search environment and comprises the same actions. On any topic these actions involve a comparison of content of the information source with information that is already known ('KKs' and 'KUs') and discovery of 'UUs'; i.e. a comparison of currently known knowledge with new information retrieved. The more matches observed, the more reliable and trustworthy the source of information becomes. It is possible that an information source contains known headlines with new detail. The new information transfers from 'KUs' into the category of 'KKs' and knowledge expands. Discovery of 'UUs' expands our knowledge further. As soon as we get information we did not know existed, this information becomes 'KUs' and presents further search options. Thus, the combination of 'KUs' and 'UUs' represents the uncertainty on the topic. Information discovery changes the level of uncertainty and its composition in an individual's knowledge by converting unknown information into known information. Following the first iteration of the algorithm, newly collected evidence will partially consist of the text that is similar to the contents of the initial knowledge base, while the major part of the new evidence will be new concepts.

Grounded theory [18, 19] has been successfully used in building a hypothesis (theory) using interviews. Grounded theory is a systematic methodology in the social sciences involving the generation of theory from data. An important characteristic of grounded theory is that it does not use any prior information, and that it builds theory only based on information that is obtained throughout the research, making it suitable in the context of evidence building with very limited prior information. Grounded theory is an integral part in our approach in order to identify the set of 'UUs' in newly gathered information through comparison of the conversion rate of 'KUs' and 'UUs' (new concepts) into 'KKs' (evidence). Total knowledge on a topic $K_{total}$ is the collection of all three sets. It is the sum of initial knowledge concepts $KK_0$, initial search objective concepts $KU_0$, while 'unknown unknowns' $UU_0$ are undefined:

$$K_{total}(0) = KK_0 + KU_0 + UU_0. \quad (3)$$

After each iteration, newly identified concepts are added to the knowledge base, thus expanding the evidence:

$$K_{total}(i) = K_{total}(i-1) + KU_i, \quad (4)$$

where 'KUs' represent new concepts on each iteration, and $KU_{(i-1)} \neq KU_{(i)}$.

Change in KU represents the conversion rate δ(KU) of new concepts in evidence and is defined as:

$$\delta(KU) = KU_i - KU_{(i-1)}. \quad (5)$$

If δ(KU) > 0, then there are still possible concepts that can be identified for evidence expansion.

If δ(KU) < 0, then we are not getting any new information and can assume that the topic is tending to exhaustion.

Conversion rate is not used to analyse the whole KU function for critical points, but to analyse the change in new concepts after each iteration. In conjunction with evidential analysis, conversion rate makes a basis for the decision on next iteration.

When $|\delta(KU)| \approx 0$ we can consider the search topic as exhausted, meaning more information will not significantly change the completeness of the evidence. The knowledge base is considered to be X% complete, if the new iteration gives X% similarity in the results with the existing knowledge base.

## IV. Applying Evidential Analysis

An effective measurement of the quality level associated with information gathered from the Web-sources is required. The Dempster-Shafer theory [20] relates to a mathematical theory of evidence and is used to express uncertain judgments of experts. In this context the hypotheses represent all the possible concepts in the knowledge base. Moreover, it is required that all hypotheses are mutually exclusive. One piece of evidence is related to a single hypothesis or a set of hypotheses. The qualitative relationship between a piece of evidence and a hypothesis corresponds to a cause-consequence chain. A piece of evidence implies a hypothesis or a set of hypotheses respectively. The strength of an evidence-hypothesis assignment, and thereby the strength of this implication, is quantified by a statement of a data source, which in our case may be a single Web-page, or the entire Google section (Books, Scholar, News, etc.).

The Dempster-Shafer theory uses a measure of basic assignment (weight of belief). This measure is correlated with an information quality measure of the Web-source. Research by Zhu & Gauch [21] presents an approach to calculate quality of a Web site on a per-topic basis by using six metrics. The following metrics are used: currency, availability, information-to-noise ratio, authority, popularity and cohesiveness. Currency is measured as the time stamp of the last modification of the document. Availability is calculated as the number of broken links on a page divided by the total numbers of links it contains. Information-to-noise ratio is computed as the total length of the tokens after pre-processing divided by the size of the document. Popularity score can be gained from the number of links pointing to a Web-page. Cohesiveness was determined by how closely related the major topics in the Web-page were. Authority of a Web-page can be measured with the equation (6), using age of domain ($age_{domain}$), number of links from other Web-sites that point to the entire domain ($N_{links}$) and size of the Web-site that relates to the amount of quality information on the Web-site ($size_{website}$):

$$Authority = \log_{10}(age_{domain} \times N_{links} \times size_{website}). \quad (6)$$

The necessary Web-site statistics can be found with an available Web-site analysis tool.

Having obtained the metrics measurements, the quality of the site was then determined by its information quality using the following equation:

$$G_i = \bar{W}_i * (a_s'' * \bar{T}_i + b_s'' * \bar{A}_i + c_s'' * \bar{I}_i + d_s'' * \bar{R}_i + e_s'' * \bar{P}_i + f_s'' * C_i), \quad (7)$$

where $\overline{W}_i$ , $\overline{T}_i$ , $\overline{A}_i$ , $\overline{I}_i$ , $\overline{R}_i$ and $\overline{P}_i$ are the means of information quantity, currency, availability, information-to-noise ratio, authority, and popularity of site *i* across topics relevant to the query. $C_i$ is the cohesiveness of site I; $a_s''$ , $b_s''$ , $c_s''$ , $d_s''$ , $e_s''$ and $f_s''$ are the weights of each quality metric.

Based on the results of evidence tests for completeness and quality, a decision is made on whether to iterate or stop. Table 1 shows all possible combinations of results for measuring evidence completeness (conversion rate) and quality.

TABLE I. CHOICE OF NEXT STEP

| Conversion rate, δ(KU) | Quality | Action |
|---|---|---|
| Positive | Positive | Continue (expand query) |
| Positive | Negative | Stop searching |
| Negative | Positive | Continue (expand query) |
| Negative | Negative | Stop (change formulation) |

Thus, the decision on the next iteration depends on the amount of new concepts coming into the knowledge base as well as the change in quality of knowledge base, if new information is to be added.

## V. EVALUATING CASE STUDY

The method is evaluated using a test topic "investing in coffee". This evaluation is an early development as a full investigation into the application of evidential analysis has not yet been completed. Therefore, the evaluation should be seen as illustrative, although the results so far are encouraging.

A textual file about the topic was randomly chosen as the initial knowledge base, and a search objective was set to "coffee producers" for the search query. The Google search engine was used for the first iteration results and received 3.5m Web-pages in a list of search results. All received pages needed to be tested for semantic closeness with respect to the text in the initial knowledge base to cancel out those Web-pages that contain keywords from the query, but are too remote in their meaning from the search topic. Starting from the first Web-page from the search results the main body text was extracted. This text is further referred to as the candidate text. Both texts (the knowledge base and candidate) were pre-processed by removing stop-words and punctuation symbols using a list of stop-words proposed by van Rijsbergen [11] for further semantic analysis. Both texts now contain only meaningful parts of speech, and are approximately 60% of their original size.

In order to compare two pieces of textual information, we first applied the Hirst & Mohammad [14] method to calculate semantic similarities between the concepts in the texts. Roget's thesaurus was chosen as a lexical resource, it contains 1044 categories of English words. We built two word-concept co-occurrence matrices, one for each text. The columns of both represented categories from Roget's thesaurus (concepts), while rows were for the words from the texts. In the first round, a word-concept co-occurrence matrix of the size (62x1044) for the knowledge base was built. The co-occurrence matrix for the candidate was of the size (152x1044). Then, having obtained the frequencies, we calculated the values of strength of association between each word and each concept in both matrices. In this experiment the order of co-occurrence was ignored. Conditional probability

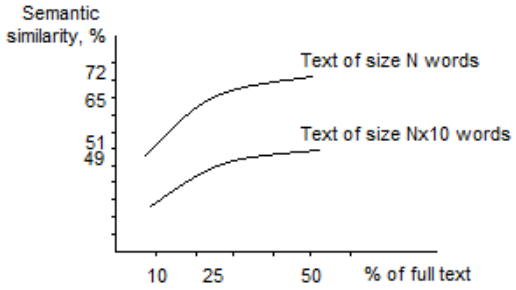


Figure3. Processing texts of different size

was chosen as a measure of strength of association between words and concepts.

Having calculated this statistics, the distributional profiles of concepts for both texts were built. For example, the word COFFEE in the thesaurus is listed under the categories FOOD, VEGETABILITY, VEGETABLE, CONDIMENT, BROWN, and REMEDY. By comparing the values of frequencies for the corresponding words and concepts, we built the distributional profiles of concepts for each of the two texts.

For example, these are the distributional profiles of concepts of the word COFFEE in both texts:

FOOD (Knowledge base): beverage(0.1429), coffee(0.2143), cup(0.2143), do(0.1429), good(0.2143), living(0.0714), …

FOOD (Candidate text): board(0.4), choice(0.2), coffee(0.2), remove(0.2), …

The values in parentheses are conditional probabilities of the words co-occurring with the concept within a window of ±5 words. Distributional profiles of concepts were treated as vectors and were compared against each other in both texts by calculating the Cosine as a measure of closeness between two probability distributions of words in concepts. We then applied the values obtained for the distance between concepts in two texts to measure the similarity between two texts. We chose a value of 0.50 as a threshold for the closeness test. Only those texts that are similar to the knowledge base with 50% or more are considered as semantically close.

It was noticed that comparing texts of different sizes results in different semantic similarity between these texts. For the experiment, two texts on the same topic were considered as initial knowledge bases. The first one has the size of approximately 150 words after pre-processing, and the second one has the size of approximately 1400 words. Each of these texts was compared against three other texts that represent 10%, 25% and 50% of the corresponding initial knowledge base text, i.e. comprise several paragraphs from the corresponding full text. Fig. 3 illustrates the dependency relation between the size of the texts and their similarity. Larger texts that comprise the same information result in a lower level of similarity between them.

In order to measure evidence completeness the Grounded theory was applied. After the first iteration every word (concept) that was not presented in the text of the initial knowledge base is considered as an 'UU' and converted to a 'KU'. While further iterating the algorithm, most of the search results repeat themselves, and are therefore, ignored. That allows us to trace new concepts more accurately and evaluate the increasing or decreasing trend of the conversion rate from 'UU' to 'KU', which is expected to start decreasing after the 5th iteration. The wider the initial knowledge, i.e. the larger the

text of the initial knowledge base, the more accurate results can be expected.

## VI. Conclusion

In this paper we presented a new framework for Web-based intelligence information acquisition and formation of a textual knowledge base. The major strength of this framework lies in the combination of existent NLP techniques, grounded theory and evidential analysis to automatically extract unknown unknowns from Web-based textual content and form a knowledge base that can be effectively manipulated by analysts to find facts (names) and associations between them (events).

The proposed similarity estimation has provided encouraging results in comparing large amounts of texts due to a higher frequency of word-concept co-occurrence, making it possible to disambiguate a sense that each word has within its context. Extracting the word sense will allow manipulation with distributional profiles of concepts that contain measures for strength of association between each word used in each of its senses (categories from the thesaurus) co-occurring with other categories, i.e. strength of association between concepts only rather than concepts and words.

The result of the experiment shows reasonable correlation between the actual meaning of the texts compared to the initial knowledge base and the calculated measures of text similarity. When two sets of texts with significant difference in size were compared, some of which were parts of the corresponding full text, the resulting similarity correlated to the size of the compared texts. The number of new concepts according to grounded theory was zero. Therefore, to achieve better accuracy one may adjust the threshold for the text similarity measure, depending on the size of the initial knowledge base. For a large text in the initial knowledge base it will be more efficient to decrease the threshold level for text similarity due to an increased number of distinct concepts involved.

Further analysis of the results shows that an intelligence knowledge base will be greatly enhanced from a richness viewpoint, if the focus of intelligence analysts is on identifying 'UUs'. A regular search of Web-based intelligence information using this new approach, especially the automated version of the grounded theory element, will yield positive results for 'UU' discovery. Future planed experimentation will be aimed at measuring the 'UU' discovery rate.

This work is in an early stage and the focus is now on incorporating evidential analysis. Detailed experiments are planned and the results of which will be published in due course.

Duplicate content is common for the Web-searches. Often the list of Web-search results contains different Web-pages with repeated content. This duplication is thought to be caused by the recent tendency of authors to paraphrase or even copy-paste the information already presented online. Therefore it is worth storing the links alongside the Web-search results to avoid repetition and compare the texts against what has already been added. This design feature will be included in the next iteration of experimentation when the Dempster-Shafer Evidential Analysis process step is included.

The next iteration of experimentation will also run with newly developed software written in Perl as opposed to this experiment which was conducted using MATLAB. The new software will be optimised for minimum running time.

## References


[1] Berners-Lee, T., "The Semantic Web". Scientific American, May 1, 2001.

[2] Rumsfeld, D., News transcript: DoD news briefing. Washington D.C.: U.S.Department of Defence,2002.

[3] Pugh, W., & Henzinger, M. (2001). *Patent No. 768947.* USA.

[4] Gomes, B., & Smith, B. (2000). *Patent No. 684542.* USA

[5] Autonomy. (2009, September 29). *Autonomy Technology Overview.* Retrieved 01 06, 2012, from Autonomy: http://publications.autonomy.com/pdfs/Power/White%20Papers/Autonomy%20Technology/20090928_PI_WP_TechOverview_web.pdf

[6] Zhou, B., Xiong, Y., & Liu, W., "Efficient Web-page main text extraction towards online news analysis". IEEE International Conference on e-Business Engineering, 2009 (ICEBE '09), (pp. 37 - 41).

[7] Adam, G., Bouras, C., & Poulopoulos, V., "CUTER: An efficient useful text extraction mechanism". Advanced Information Networking and Applications Workshops (WAINA), 2009, pp. 703-708. Institute of Electrical and Electronics Engineers ( IEEE ).

[8] Hu, G., & Zhao, Q., "Study to eliminating noisy information in Web-pages based on data mining". Sixth International Conference on Natural Computation (ICNC 2010), Volume 2, pp. 660 - 663.

[9] Fu, L., Meng, Y., Xia, Y., & Yu, H., "Web-content extraction based on Web-page layout analysis". Second International Conference on Information Technology and Computer Science (ITCS 2010), Ukraine, pp. 40 - 43.

[10] Yi, L., Liu, B., & Li, X., "Eliminating noisy information in Web-pages for data mining". Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, 2009, New York, NY, USA: ACM, pp. 296 – 305.

[11] van Rijsbergen, C. J., "Information retrieval". London: Butterworths, 1979.

[12] Luhn, H., "The automatic creation of literature abstracts". IBM Journal, 1958, pp. 159 – 165.

[13] Fox, C. J., "A stop list for general text". ACM Special Interest Group on Information Retrieval Forum 24, 1990, pp. 19 – 35.

[14] Hirst, G. & Mohammad, S., "Measuring semantic distance, using distributional profiles of concepts". New York: Association for Computational Linguistics, 2006.

[15] Fellbaum, C., "WordNet: an electronic lexical database". Cambridge, MA, USA: The MIT Press, 1998.

[16] Rubenstein, H., & Goodenough, J., "Contextual correlates of synonymy". Communications of the ACM , 8 (10), October, 1965, pp. 627 – 633.

[17] Corley, C., & Mihalcea, R., "Measuring the semantic similarity of texts". EMSEE '05 Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, 2005, pp. 13 – 18. Stroudsburg, PA, USA: Association for Computational Linguistics

[18] Martin, P., & Turner, B., "Grounded theory and organizational research". The Journal of Applied Behavioral Science , 22 (2), 1986, pp.141 – 157.

[19] Corbin, J., & Strauss, A., "Basics of qualitative research: techniques and procedures for developing grounded theory" (3rd edition ed.). London: Sage Publications, 2008.

[20] Shafer, G., "A mathematical theory of evidence". Princeton: Princeton University Press, 1976.

[21] Zhu, X., & Gauch, S., "Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web". SIGIR '00 Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, 2000, pp. 288 – 295. ACM New York, NY, USA.