Subject domain models creation on the basis of the Internet content monitoring

Dmitry V. LANDE, Doctor of Sciences, Head of Department of Specialized Modeling Tools, Professor of National Technical University of Ukraine “I. Sikorsky Kyiv Polytechnic Institute”

Kharkiv, 2017

National Academy of Sciences of Ukraine
Institut for Information Recording
Internet has a huge amount of information, which could be helpful in solution for various problems and targets. It is important to make choice which is exact available to the user interest. A significant amount of information resources in global networks contains various expert assessments, some of which is related to carrying out information impacts and operations. Registration and classification of information sources is an important base for finding development directions in economy, science, technology, etc. The same, as for problem solving for security of human, society and state. Every specific case has important tool of analysis - the subject domain model - the ontology.
It will be represent theoretical and technological bases for creation the subject domain models. Models can be used for creation of scenarios of information support of decision-making are provided. At the same time knowledge sources are considered the text data massifs and resources of the modern scientometric and encyclopedic services.
Theoretical and technological solutions

- Finding messages on topics of interest in Internet.
- Tracking information flows (stories), relevant topics, events, and processes.
- Determination of the dynamics of information flows.
- Definition of abnormal and critical point in the dynamics of the thematic information flows.
- Identify major events and objects of thematic information flow.
- Creating of subject domain ontology.
- Visualize relationships of monitoring objects.
- Forecasting developments.
A system of multilingual full-text databases has been created, using a universal encoding system (UTF). The implementation of this system provides the ability to quickly locate local information from different countries, monitor information plots from different points of view, analyze the impact of individual events on different segments of the population.
Monitoring system for multilingual full-text databases

<table>
<thead>
<tr>
<th>No.</th>
<th>Platform</th>
<th>Status</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Youtube</td>
<td></td>
<td>Sources</td>
</tr>
<tr>
<td>2.</td>
<td>Twitter</td>
<td></td>
<td>Sources</td>
</tr>
<tr>
<td>3.</td>
<td>Weibo.com</td>
<td></td>
<td>Sources</td>
</tr>
<tr>
<td>4.</td>
<td>Odnoklassniki</td>
<td></td>
<td>Sources</td>
</tr>
<tr>
<td>5.</td>
<td>LiveJournal</td>
<td></td>
<td>Sources</td>
</tr>
<tr>
<td>6.</td>
<td>LiveInternet</td>
<td></td>
<td>Sources</td>
</tr>
</tbody>
</table>

**Запрос: 中国**

1. **中国神舟十一号飞船成功发射** HD 2016.10.17
   - 中国神舟十一号飞船成功发射 HD 2016.10.17 神舟、神舟11、神舟十一号、天宮二号、太空船、航天飛船、飞船

YouTube: RyukyuSARs 2016.10.21 14:30
https://www.youtube.com/watch?v=sBu8KJXYh0
UCNV EjajT4i9VN1HH_vPnQ
Proxy:
../proxy/yt/20161021/The Thanh.txt
Thematic information streams can be associated with time series (the intensity of publications per unit of time), for the analysis of which the formal methods are increasingly used: statistical, fractal, Fourier or wavelet analysis. The analysis of these numerical series allows to reveal trends, cycles, anomalies, correlations, etc.
With the help of methods for extracting data from texts that are used in the analysis of thematic information flows, it is possible to form networks of interrelations of concepts whose nodes are reference words, names of persons, toponyms, brands, organizations, etc. Analysis of these networks makes it possible to identify explicit and implicit connections between individual concepts, the weight of concepts, refine the criteria for the formation of the information flow, and organize "hints" in the interfaces of intelligent systems.
Keywords for searching in the text are chosen taking into account such a word property as "discriminant force", for the definition of which there are three classical methods - TFIDF, dispersion, entropy. New network methods are also used.
For each term from a textual array, the value of TFIDF is equal to the product of the frequency of this word in the text fragment (Term Frequency) by the logarithm of the inverse of the number of separate fragments of text in which this word met (Inverse Document Frequency).

For each word $i$ in the text array, which consists of $N$ documents, the number of $df\ (i)$ documents containing this word is counted, as well as the total frequency of occurrence of this word in the text array - $n\ (i)$. After that, the average TFIDF for each word is calculated:

$$\text{tfidf}(i) = \frac{n(i)}{N} \log \left( \frac{N}{df(i)} \right)$$
TFIDF gives not quite correct results on text arrays from documents with different sizes. Therefore, its modification - Okapi BM25:

\[
\text{TFIDF}(w_i) = \text{IDF}(w_i) \times \frac{f(w_i, D) \times (k_1 + 1)}{f(w_i, D) + k_1 \times (1 - b + b \times \frac{|D|}{\text{avgdl}})}
\]

where \(f(w_i, D)\) - frequency of a word in a document \(D\); \(|D|\) - size of document \(D\); \(\text{avgdl}\) - average document length in the collection; \(k_1\) and \(b\) - parameters. IDF calculated by the formula:

\[
\text{IDF}(w_i) = \log \frac{N - n(w_i) + 0.5}{n(w_i) + 0.5},
\]

where \(N\) - total number of documents in the array, \(n(w_i)\) - number of documents containing the term \(w_i\).
If we consider terms as events, we can find the so-called entropic estimate of the weight of the term (Pointwise Mutual Information, PMI):

\[
m(w) = -\log_2 \frac{p(w \mid d)}{p(w)} = -\log_2 \frac{n_{w,d}}{n_w},
\]

where: \( n_{w,d} \) - the number of entries of the term in the text of the document, \( n_w \) - total number of entrances.
As a function that matches a word from a text, we can consider different weighted estimates, in particular, statistical dispersion ones. Dispersion $\sigma_A$ estimate for some word $A$ from the text is calculated as

$$\sigma_A = \sqrt{\langle \Delta A^2 \rangle - \langle \Delta A \rangle^2} \frac{1}{\langle \Delta A \rangle},$$

where: $\langle \Delta A \rangle$ - the average distance (in words) between the occurrences of the word $A$ in the test; $\langle \Delta A^2 \rangle$ - the mean square of the distance between the occurrences of the word $A$ in the text.
The first step for text analysis is representation of this text as a aggregation of nodes and links, creation of Language Network. There is meaning of application of the theory of complex networks.

Nodes can be linked together, if the words corresponding to them are nearby in the text, belong to one sentence, syntactically or semantically.
**L-space.** Associate neighboring words that are included in one sentence. The number of neighbors for each word (word window) is determined by the interaction radius $R$. The case $R = 1$ is most often considered.

**B-space.** Bigraph. Consider nodes of two types, corresponding to the sentence and words that are included in them.

**P-space.** All the words that are included in one sentence are linked together.

**C-space.** Proposals are linked together if they use the same words.
From the selected terms (unigram, bigram, trigram), networks of natural hierarchies of terms are constructed. As nodes are considered terms. Links correspond to the entry of some terms into others.
Within the framework of Digital Signal Processing and Complex Network theories, several methods of constructing networks based on numerical series are proposed.

Among them - a family of methods for constructing Visibility Graph, in particular, the so-called Horizontal Visibility Graph (HVG).

These approaches also allow building network structures based on texts. In these cases, certain initial numerical weight values are assigned to certain words or phrases in a certain special way.
Method of horizontal visibility graph forming
Compactification of horizontal visibility graph
As an example of the original text corpus, annotations of electronic preprints Arxiv (http://arxiv.org) on the subject of information retrieval (heading cs.IR, more than 500 documents for 5 years) are considered.

**Example 1: Source corpora arxiv.org**

**Information Retrieval**

**Authors and titles for recent submissions**

- Fri, 5 Sep 2014
- Thu, 4 Sep 2014
- Wed, 3 Sep 2014
- Tue, 2 Sep 2014
- Mon, 1 Sep 2014

[ total of 13 entries: 1-13 ]
[ showing up to 25 entries per page: fewer | more ]

Fri, 5 Sep 2014


On the Accuracy of Hyper-local Geotagging of Social Media Content

David Flatow, Mor Naaman, Ke Eddie Xie, Yana Volkovich, Yaron Kanza

Comments: 10 pages
Subjects: Information Retrieval (cs.IR); Social and Information Networks (cs.SI)
Example 1: Source corpora arxiv.org – NNHT visualization
With the help of the algorithm HITS provides the choice of the best **authors** (nodes, which will be linked to) from the information array and **hubs** (nodes from which the links go). The term is a good hub, if it refers to important phrases, and vice versa, a good author, if it is referred to by important authors. In accordance with the HITS algorithm, for each node of the network $v_j$, is recursively calculated as the author of $a(v_j)$ and the hub $h(v_j)$ according to the formulas:

$$a(v_j) = \sum_i h(v_i); \quad h(v_j) = \sum_i a(v_i).$$
### Example 1: Terms - the best authors / hubs

<table>
<thead>
<tr>
<th>Authors</th>
<th>Hubs</th>
</tr>
</thead>
<tbody>
<tr>
<td>WEB_SEARCH_ENGINE</td>
<td>SEARCH</td>
</tr>
<tr>
<td>SEARCH_ENGINE_RESULT</td>
<td>ENGINE</td>
</tr>
<tr>
<td>BASED_SEARCH_ENGINE</td>
<td>SEARCH_ENGINE</td>
</tr>
<tr>
<td>VERTICAL_SEARCH_ENGINE</td>
<td>WEB</td>
</tr>
<tr>
<td>SEARCH_ENGINE_COMPANIES</td>
<td>RESULT</td>
</tr>
<tr>
<td>GEOGRAPHIC_SEARCH_ENGINE</td>
<td>SEARCH_RESULT</td>
</tr>
<tr>
<td>ENTITY_SEARCH_ENGINE</td>
<td>WEB_SEARCH</td>
</tr>
<tr>
<td>SEARCH_RESULT_CLUSTERING</td>
<td>IMAGE</td>
</tr>
<tr>
<td>TAG_SEARCH_RESULT</td>
<td>PAGE</td>
</tr>
<tr>
<td>SEARCH_RESULT_PRESENTATION</td>
<td>CLUSTERING</td>
</tr>
<tr>
<td>PERSONALIZE_SEARCH_RESULT</td>
<td>TAG</td>
</tr>
</tbody>
</table>
The NNHT as network can be considered as the basis for the formation of other links between its nodes. If we denote the incidence matrix of NNHT by the letter $A$, then the matrices $AA^T$ and $A^TA$ will reflect the relations of occurrences of these types:

Example 1: Associative connections

![Diagram showing associative connections]

- $a_i \sim a_j$
- $a_k \sim a_i \sim a_j$
- $\delta$
Example 1: Fragment of NNHT with associative links
Example 1: Conclusions on the NNHT algorithm

An algorithm for constructing networks of natural hierarchies of term (NNHT) based on text analysis is proposed.

An algorithm for constructing associative relationships between terms in NNTH is proposed.

The use of the HITS algorithm for selecting the most important elements in the NNTP is proposed.

The method of visualization of NNTP fragments is proposed.

A language network built using the proposed methodology can be used as a basis for constructing an ontology of the domain, as a ready-to-use navigation tool in information arrays, and for organizing contextual prompts for users of intelligent systems.
Using the content monitoring system, we investigate the problem (for example, Brexit). To do this, a request is made to the content monitoring system and the activity period is determined.
Documents come from the content-monitoring system of the global network. From the documents the concepts and terms necessary for constructing the domain model are extracted.
Example 2: Network of natural hierarchies of terms
Example 2: Fragment of of natural hierarchies of terms
Example 2: Nodes of NNHT with the highest PageRank

<table>
<thead>
<tr>
<th>Id</th>
<th>PageRank</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRITISH_Government</td>
<td>0.035354</td>
</tr>
<tr>
<td>IRISH_INTEREST</td>
<td>0.035354</td>
</tr>
<tr>
<td>STUDENT_EXCHANGE_PROGRAMMES</td>
<td>0.021389</td>
</tr>
<tr>
<td>DONALD_TRUMP_QUESTION</td>
<td>0.01848</td>
</tr>
<tr>
<td>ANTI_BREXIT_PROTEST</td>
<td>0.016767</td>
</tr>
<tr>
<td>MEP_NI_SECRETARY</td>
<td>0.016087</td>
</tr>
<tr>
<td>CHARLIE_FLANAGAN_GOOD</td>
<td>0.014711</td>
</tr>
<tr>
<td>CLINTON_SEIZES</td>
<td>0.014399</td>
</tr>
<tr>
<td>IRELAND_UK_BOND</td>
<td>0.014399</td>
</tr>
<tr>
<td>VOTE_BANK_BAILOUTS</td>
<td>0.01406</td>
</tr>
<tr>
<td>BREXIT_IMPACT_FADES</td>
<td>0.014038</td>
</tr>
<tr>
<td>POST_BREXIT_REBOUND</td>
<td>0.012848</td>
</tr>
<tr>
<td>SF_MEP_NI</td>
<td>0.011544</td>
</tr>
<tr>
<td>LORETTA_LYNCH_AFTER</td>
<td>0.011544</td>
</tr>
<tr>
<td>BREXIT_SOCIAL_DEMOCRATIC</td>
<td>0.011407</td>
</tr>
<tr>
<td>AS_SF_MEP</td>
<td>0.011158</td>
</tr>
<tr>
<td>IMPACT_ON_SOUTHEAST</td>
<td>0.010622</td>
</tr>
<tr>
<td>POST_BREXIT_SPIKE</td>
<td>0.010052</td>
</tr>
<tr>
<td>SURVIVE_BREXIT_THERESA</td>
<td>0.010052</td>
</tr>
<tr>
<td>BREXIT_VOTE_GOOD</td>
<td>0.009967</td>
</tr>
<tr>
<td>IRELAND_BUT_BREXIT</td>
<td>0.009967</td>
</tr>
<tr>
<td>BREXIT_VOTE_HOUR</td>
<td>0.009967</td>
</tr>
<tr>
<td>ASSESSES_BREXIT</td>
<td>0.009967</td>
</tr>
<tr>
<td>BREXIT_CAMPAIGNER</td>
<td>0.009967</td>
</tr>
</tbody>
</table>
**Example 2: Nodes of NNHT with the highest hubs values (HITS)**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>369.30143</td>
<td>BREXIT</td>
</tr>
<tr>
<td>2</td>
<td>26.36127</td>
<td>SOCIAL</td>
</tr>
<tr>
<td>3</td>
<td>17.47190</td>
<td>IMPACT</td>
</tr>
<tr>
<td>4</td>
<td>17.09261</td>
<td>BREXIT_SOCIAL</td>
</tr>
<tr>
<td>5</td>
<td>16.97070</td>
<td>PROTEST</td>
</tr>
<tr>
<td>6</td>
<td>16.68476</td>
<td>REBOUND</td>
</tr>
<tr>
<td>7</td>
<td>8.97908</td>
<td>SOCIAL_DEMOCRATIC</td>
</tr>
<tr>
<td>8</td>
<td>8.68602</td>
<td>LYNCH</td>
</tr>
<tr>
<td>9</td>
<td>8.68508</td>
<td>IMPACT_FADES</td>
</tr>
<tr>
<td>10</td>
<td>8.40086</td>
<td>BREXIT_IMPACT</td>
</tr>
<tr>
<td>11</td>
<td>8.39978</td>
<td>LYNCH_AFTER</td>
</tr>
<tr>
<td>12</td>
<td>8.39214</td>
<td>BREXIT_PROTEST</td>
</tr>
<tr>
<td>13</td>
<td>8.39214</td>
<td>ANTI_BREXIT</td>
</tr>
<tr>
<td>14</td>
<td>8.29681</td>
<td>BREXIT_REBOUND</td>
</tr>
<tr>
<td>15</td>
<td>8.29579</td>
<td>REFERENDUM_BRITAIN</td>
</tr>
<tr>
<td>16</td>
<td>8.20576</td>
<td>AS_SF</td>
</tr>
<tr>
<td>17</td>
<td>8.20365</td>
<td>HAMPER</td>
</tr>
<tr>
<td>18</td>
<td>8.11353</td>
<td>VOTE_HOUR</td>
</tr>
<tr>
<td>19</td>
<td>8.11353</td>
<td>VOTE_GOOD</td>
</tr>
<tr>
<td>20</td>
<td>8.11353</td>
<td>SURVIVE_BREXIT</td>
</tr>
<tr>
<td>21</td>
<td>8.11353</td>
<td>IRELAND_BUT</td>
</tr>
<tr>
<td>22</td>
<td>8.11353</td>
<td>DIVORCE</td>
</tr>
<tr>
<td>23</td>
<td>8.11353</td>
<td>CAMPAIGNER</td>
</tr>
<tr>
<td>24</td>
<td>8.11353</td>
<td>BREXIT_SPIKE</td>
</tr>
<tr>
<td>25</td>
<td>8.11353</td>
<td>ASSESSES</td>
</tr>
</tbody>
</table>
Example 3: Building a domain model based on Google Scholar Citations
Fragment of the interface

Google

label:comparative_law

Alec Stone Sweet
Professor of Law and Political Science, Yale University
Verified email at yale.edu
Cited by 8722
Comparative law, comparative politics, international law, international relations, European integration

Nuno Garoupa
Professor of Law
Verified email at illinois.edu
Cited by 2752
Law and Economics, Comparative Law

Annelise Riles
Professor of Law and Professor of Anthropology, Cornell University
Verified email at cornell.edu
Cited by 2297
Comparative law, global financial markets regulation, Asia, human rights
Example 3: Building a domain model based on Google Scholar Citations. Sounding Algorithm

1. The short list of the base tags, defined in an expert way.
2. One tag is chosen from the list.
3. The performed search represents the web-page corresponding with the chosen tag.
4. The neighboring tags, comprised in the page, are added to the forming network.
5. From the neighboring tags are chosen the ones, pages of which are planned to be proceed for the further analysis. This tag is the one with the highest rate, which responds the theme of the chosen subject area, and the transition to which hasn’t been executed.
6. If such a tag been chosen the jump to the step 3 is to be provided.
7. If there is no such a tag, and the list of base tags is not fulfilled, then the segue to the next base tag from the initial tag list is provided (Step2). Otherwise the network is considered to be built.
Example 3: Building a domain model based on Google Scholar Citations.

Principle. Example
Example 3: Building a domain model based on Google Scholar Citations. Fragment of the concepts network.
Example 3: Building a domain model based on Google Scholar Citations. Some conclusions

1. In the proposed domain model, ontological links are used to link the sectors of interest of individual scientists. In fact, some compactification of the biographer "scientist - the branches of science that interest him" is actually considered. These connections suggest a semantic connection.

2. The approach of formation of the domain model is proposed and implemented. In this case, the knowledge pre-determined by the participants in the Google Scholar Citations project is applied.

3. The model can be applied to various branches of science.
1. On the main national Wikipedia page in the search line the initial word is given, eg (for English version - «Albert Einstein», for Chinese one - «阿尔伯特·爱因斯坦», etc.)

2. The search window opens. It contains information about concept, according to the task on the Step1. The initial word/word combination is a graph vertex, which will be formed as the result of scanning.

3. All terms-concepts corresponding the hyperlinks on the chosen page, are added to the formed graph. All the words/words combinations are the nodes of the graph. The edges to them are formed from the initial node.

4. The next transition is made by the first not involved hyperlink from the examining pages.

5. In text on the page to which the transition has been made the search of shortened researcher’s name (eg, Einstein, 爱因斯坦) or keyword (eg, physics, relativity, 物理学, 相对性) is to be carried out.

6. In case, if there is a shortened researcher’s name or keyword is found, the transition to the Step 4 is made and accordingly from the node – word/word combination of the current search the new nodes are built.

7. If there is no word/word combination in the text – the given graph branch is considered to be built.

8. The next transition presumes pass to the page, which had been scanned – the word is not added as a graph node, and the feedback to the created node is formed.

9. All the operations under steps 4-9 repeat until the not involved hyperlinks, chosen from the page, are left. In another case the graph is considered to be built.
Example 4: Running the program

Bibliometrics

From Wikipedia, the free encyclopedia

Bibliometrics is statistical analysis of written publications, such as books or articles.[1] Bibliometric methods are frequently used in the field of library and information science, including scientometrics. For instance, bibliometrics are used to provide quantitative analysis of academic literature[2] or for evaluating budgetary spending.[3] Citation analysis is a commonly used bibliometric method which is based on constructing the citation graph, a network or graph representation of the citations between documents. Many research fields use bibliometric methods to explore the impact of their field,[4] the impact of a set of researchers, or the impact of a particular paper. Bibliometrics also has a wide range of other applications, such as in descriptive linguistics, the development of thesauri, and evaluation of reader usage.

Information science

General aspects
- Information access
- Information architecture
- Information management
- Information retrieval
- Information seeking
- Information society
- Knowledge organization
- Philosophy of information
- Science, technology, and society

Related fields and sub-fields
- Bibliometrics
- Categorization
- Classification
- Computer data storage
- Cultural studies
- Data modeling
- Informatics
- Information technology
- Intellectual freedom
- Intellectual property - Memory
- Library and information science
- Preservation - Privacy
- Quantum information science

Step: 1; Terms: 14

2: Information science

- Data science
- Library science
- Informatics

- Information retrieval
- Information access
- Information architecture
- Information management
Example 4: Fragment of network
### Example 4: Ranking of concepts

<table>
<thead>
<tr>
<th>Concept</th>
<th>Node Degree</th>
<th>PageRank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Information_science</td>
<td>19</td>
<td>0.061</td>
</tr>
<tr>
<td>Information_technology</td>
<td>19</td>
<td>0.061</td>
</tr>
<tr>
<td>Information_society</td>
<td>19</td>
<td>0.061</td>
</tr>
<tr>
<td>Information_management</td>
<td>19</td>
<td>0.061</td>
</tr>
<tr>
<td>Information_retrieva</td>
<td>18</td>
<td>0.055</td>
</tr>
<tr>
<td>Information_architectur</td>
<td>18</td>
<td>0.055</td>
</tr>
<tr>
<td>Information_acces</td>
<td>17</td>
<td>0.052</td>
</tr>
<tr>
<td>Bibliometrics</td>
<td>19</td>
<td>0.049</td>
</tr>
<tr>
<td>Categorization</td>
<td>17</td>
<td>0.049</td>
</tr>
<tr>
<td>Information_seeking</td>
<td>17</td>
<td>0.049</td>
</tr>
</tbody>
</table>
By results of researches in the considered direction 17 monographs and more than 100 articles are published.


Some publications of the author

http://dwl.kiev.ua


Thank you for your attention!

Dr. of Sci. D. Lande
http://dwl.kiev.ua