The present day analysis strives to extract intelligence from complex, often unstructured data (text, video, and so on) in real time from multiple sources.

For more than 10 years, «I-Teco» has been developing knowledge extraction technology and analytical processing of textual information. I-Teco has developed a system for knowledge extraction from documents «Analytical courier», a tool for the analysis of information. Its main features:

- analysis of stakeholder views regarding different objects (Opinion Mining);
- entity-level sentiment analysis (N-view Fine-grained Sentiment Analysis);
- relation analysis between objects in a collection of documents;
- identification of sources of information campaigns and attacks, search for duplicates;
- topic tracking in the flow of incoming documents;
- sociometry of authors and messages, geolocation of the authors;
- event tracking, the frequency analysis of topics, social media communications, etc.;
- assessing the probability of events of required types and their significance;
- thematic classification of the incoming messages;
- author and profiles ranking of social media posts;
- identification sociologically significant features of Author profiles;
- analysis of the social graph, identification and analysis of online communities.

«Analytical Courier» can be used to identify sources of information campaigns and attacks and to predict the occurrence of certain events to reveal their information profiles, etc.

The system includes a fault-tolerant, scalable Big Data repository, allowing to safely and efficiently handle billions of documents. It can be used separately from the system, and is a part of many applications designed for different users. It can remove / unload documents using rules: for example, it can remove documents given the «age» of the selected sources. This allows efficient use of disk storage space.

The system supports English and Russian in the default configuration. Other languages, including Farsi, Urdu, Chinese and many others can be added by request.
System’s features

The system uses modern methods and techniques of data analysis.

Data sources:
- Web monitoring via different search engines;
- targeted monitoring with Internet robots for all types of sources. Users may specify the sites for monitoring. The basic sources for continuous monitoring include:
  - more than 4,000 Internet sites;
  - social networks: VKontakte, Facebook, Google+, Classmates, My World, etc.;
  - blogs: LiveJournal, Blogspot, LiveInternet, Blog.ru, etc.;
  - microblogs: Twitter;
  - various forums;
  - review and opinion sites about a variety of products, services and organizations;
- automatic author profile loading and the formation of storage profiles;
- calculation of page rating and authors’ ranking;
- monitoring of e-mail messages, documents from the network file system and database records;

Loading the repository:
- page cleaning, text similarity analysis, document meta-information detection, continuous loading into the message repository;
- real-time entity extraction and their relationship identification, opinion mining and sentiment analysis in several languages;
- documents’ theme classification;
- event tracking with user alert;

Search for information:
- Personalized Search (SERPs ranking of documents, taking into account the history of user queries);
- multilingual and cross-language information retrieval;
- multilingual semantic search using modern thesauri of Russian and other languages;
- Semantic search with natural language queries; search for documents with similar content, search of reprints;
- Refined search found documents;

Search Results:
- documents clustering for duplicates, information sources, and other document properties;
- search results with their thematic profile by means of frequency distribution of authors, key subjects and themes, sources, publication dates and other document properties;
- semantic clustering of results by means of synsets. For example, in the frequency distribution of different themes comprising a single synset it will be provided by a single value - the head word (for example, Russian President Vladimir Putin -> President of Russia, President Putin, Vladimir Vladimirovich. Putin ...). The same is true for semantic and cluster maps, key themes, etc.. A user can add new synonyms to a synset to improve coverage.

Analytical processing:
- implementation of complex scenarios for analytical processing of search results. For example, a user can analyze possible topical (cluster) patterns in a document collection, and then build a relation map of themes for the selected cluster, and then perform object-oriented sentiment analysis (displayed along with frequency distribution), etc.;
• thematic summaries for collections of documents;
• categorization, automatic selection of the thematic groups in the SERP documents (cluster analysis) and thematic abstracting of documents;
• identification of key documents in the SERPs, creation of tag clouds for collections of documents;
• visualization of the relation map of themes in the collection of documents in the form of an interactive semantic map, link analysis (also possible to get the corresponding documents);
• link analysis of objected defined by the user in the whole collection of documents;
• construction of an interactive multi-dimensional frequency distributions of statements, their sentiment or message properties (i.e. sources, authors, ...);
• trend analysis;

Knowledge Management:
• Entity Extraction (People, Companies, Places, Dates, Brands, Events, etc) and Identification of Relations between Entities. Coreference and anaphora resolution to improve coverage;
• Possibility of edit synsets «on the fly» with simultaneous change of all corresponding results including semantic search. This unique feature allows the analyst to independently configure and use his/her conceptual space based on inclusion of the theme or object encountered in a text into an existing synset;
• thesauri management system.

Software architecture
The system can be operated using cloud computing. The server part of the application, accessible via a web interface, is implemented in Microsoft.NET. Some functions are delegated to client computers, which increases server performance greatly.

The system makes use of an easy-to-configure distributed database containing documents and their semantic representation, full-text, attribute, and semantic indexes. It is implemented on different software platforms offering the best price-quality-development ratio and is optimized for use in Big Data projects. This storage system can handle a virtually unlimited number of documents without performance loss. The performance is optimized by means of distributed caching of large data sets, following trendy cash Memory DBMS principle, as well as by server load balancing. The server line make use of both horizontal and vertical scaling as well as data mirroring.

Competitive advantages

«Analytical Courier» is an intelligence analysis system making use of different analytical methods including unique opinion mining and sentiment analysis algorithms. All the knowledge extracted by the system is aggregated in real time. Unlike its competitors, «Analytical Courier» allows users to create and edit data samples, which immediately influence the system output. The system provides framework for a professional search engine combining the knowledge of all its users.

Some other competitive advantages:
• The information retrieval process is fast and its output are of a significantly higher quality comparing to those of competitive systems.
• Synsets can be created automatically or “on the fly” by the user. All topics shown in the system output are represented by the dominant term of the synset they belong to.
• The systems allows emotive sentences extraction distinguishing between the subject and the object of an utterance.

<table>
<thead>
<tr>
<th>Объект</th>
<th>Субъект</th>
<th>Выполняемое (360)</th>
<th>Точность</th>
<th>Автор</th>
<th>Источник</th>
<th>Дата публикации</th>
</tr>
</thead>
<tbody>
<tr>
<td>БОВИН КРИТИЧЕСКИЙ ИССЛЕДОВАНИЯ ГОСУДАРСТВЕННОЙ ЯЗЫКИ И ЛИТЕРАТУРЫ</td>
<td>ИТАР-ТАСС</td>
<td>США заявили, что они изменили выводы, касающиеся деятельности Российской Федерации и Европейского союза.</td>
<td>Роман Кудрявцев</td>
<td>ria.ru</td>
<td>13.06.2014</td>
<td></td>
</tr>
<tr>
<td>ПУЛЬОВ СОДЕРЖИМАЯ СИСТЕМА-СЕСА</td>
<td>Иван Лукичев</td>
<td>Лукичев в интервью с журналом &quot;Смена&quot;, говоря о важности права в рамках Европейского союза, заявил о значимости международного сотрудничества в борьбе с коррупцией.</td>
<td>rts.ru</td>
<td>rts.ru</td>
<td>16.06.2014</td>
<td></td>
</tr>
<tr>
<td>ПРЕДСЕДАТЕЛЬ КОМИССИИ ПО РАЗВИТИЮ ПРОТЕКАНСКИХ / ПОЛИТИЧЕСКИЕ ПРОЦЕССЫ</td>
<td>Вторник</td>
<td>Вторник в интервью с журналом &quot;Смена&quot;, говоря о важности прав в рамках Европейского союза, заявил о значимости международного сотрудничества в борьбе с коррупцией.</td>
<td>zmms.ru</td>
<td>zmms.ru</td>
<td>15.06.2014</td>
<td></td>
</tr>
<tr>
<td>ОНУГАЙ</td>
<td>ПРЕДСЕДАТЕЛЬ ПРАВИТЕЛЬСТВА ФИНАНСОВ КЛАССС</td>
<td>&quot;Смена&quot;</td>
<td>Вторник в интервью с журналом &quot;Смена&quot;, говоря о важности прав в рамках Европейского союза, заявил о значимости международного сотрудничества в борьбе с коррупцией.</td>
<td>rts.ru</td>
<td>rts.ru</td>
<td>16.06.2014</td>
</tr>
<tr>
<td>НОВАЯ ПОСЕЛА</td>
<td>НОВАЯ ПОСЕЛА</td>
<td>Высокий платежеспособный граждан, проживающий в Лисовке, оказался на 30% ниже среднего в областях.</td>
<td>rts.ru</td>
<td>rts.ru</td>
<td>16.06.2014</td>
<td></td>
</tr>
<tr>
<td>ПОЛЯНСКИЙ</td>
<td>ПОЛЯНСКИЙ</td>
<td>Президент, который проводит в Польше.</td>
<td>rts.ru</td>
<td>rts.ru</td>
<td>16.06.2014</td>
<td></td>
</tr>
</tbody>
</table>

• The system allows on-line analysis of virtually any amount of data, with the processing in most cases taking only a few seconds.

Demos

YouTube demo-videos (Part 1, Part 2, Part 3, Part 4) demonstrate the basic features of the "Analytical Courier".

Starting Part 1

The video contains the following information:
• General information about the system;
• Document retrieval model;
• Review of analytical tools for document processing,

Demonstrates the following tools:
• Topic-based profile of the collection of documents;
• Cluster map;
• Semantic map.

Starting Part 2

The video shows tools for analytical processing of documents:
• A list of sentiment-related phrases found in documents;
• Statistical distribution of documents;
• Mapping of the geographical entities found in the documents.

Starting Part 3

The video shows tools for analytical processing of documents:
• Social graph;
• Thematic Index;
• Dashboards, and the procedure for reporting.
Part 4

The video contains information about the additional capabilities of the system:

- Description of the semantic types;
- User settings;
- Administrating the system;
- Built-in help,
as well as information about the development of the system.

Examples of use

Analytical and security departments of banks

Analysis of the creditworthiness of the client, identification of the objects that make suspicious payments, detection of dissemination of confidential information, etc.

Insurance companies

Identification of fraud, unscrupulous insurers, their links, as well as patterns of client-related events of the insurance company.

The analytical departments of production companies

Analysis of the most frequent problems in the production, tracking the market sentiment towards the quality of goods.

Marketing departments and companies

Providing decision-makers with the information to develop the optimal solution to the problems they face.

Special services, law enforcement authorities

Monitoring of events, objects, issues and analysis of the relationships between the entities.

Cloud-based information services

Monitoring of events, objects, and problems can be performed using the cloud-based system deployed in the data the Company data-center. The user can specify the sources of the information, or the subject of downloadable documents. Daily «the cloud» loads several million messages from the most relevant sources.
Demonstration of the system components

Sentiment analysis

Detecting opinions in the text, also known as sentiment analysis is used to determine the emotional attitude to something or someone, for example, to products, brands, to assess people's attitude to the events in the political life of the country, etc.

To test a component on your example and learn more about it, you can here.

Opinion mining

Opinion mining finds the opinions that authors express in the form of direct or indirect speech. Opinion mining can be used to extract statements made by authors, regarding a predefined or an arbitrary object. If a text contains any object, the system can find what sentiment is used to express opinion towards this object.

For more information about the work of the component, and you can test it here.

Named entity recognition

Automatic recognition of named entities used to find proper names, terms, company names, etc. The system uses statistical language models. The number of types and their description is given during training and is independent of the dictionary.

For more information about the work of the component, and you can test it here.

Send a request for information to: ak@i-teco.ru

Related Products

«X-Files» — the system of formation and analysis of the dossier on objects monitoring.