

Extrakce informací z úplných textů

Jan Žbirka, Tovek Partner

INFORUM 2006

Praha, 23.5.2006

Zpracování úplných textů

- Fulltextové vyhledávání (3. generace)
- Vizualní analýzy
- **Extrakce informací**

Podstata extrakce informací

- nahrazování určitých slov (výrazů)
- „nálepkami“ „pojmu“
- odkazování na nálepky pojmů ve vyšších vrstvách
- získávání strukturovaných faktů
- za cenu ztráty (nepodstatného?!) textu

Extrakce informací z textů

The screenshot displays a software interface for text extraction. At the top, the 'Input file content' section shows the text: 'Russia Sold \$6Bln Worth of Arms in 2005' followed by a paragraph: 'In 2005 Russia sold \$6 billion worth of arms, which is an increase of \$500 million on the previous year. Along with oil and gas, military products remain one of country's main exports.' An 'Extract' button is located to the right of this section.

The 'Extraction result' section on the left shows a hierarchical tree structure of extracted entities. The root node is 'securMilitary', which branches into several categories, each with a blue circular icon and a plus sign. The categories include: '/locationArms/Russia sell \$6Bln Worth of arr', '/locationName/Russia/Russia', '/businessOper/sell', '/dollarNum/\$6Bln', and '/weaponsN/Worth of arm'. Other categories include '/locationArms/Russia sell \$6 billion worth of', '/business/increase', '/dollarComplex/\$500 million', '/militaryBusiness/military product', '/business/export', '/functionComplex/Russian defense Minister', '/functionComplex/President|president Vladii', '/business/order', '/dollarComplex/\$22 billion', '/business/industry', '/business/order', '/militaryBusiness/arm industry', '/location/Russia', and '/business/plan'.

The main text area on the right shows the original text with various entities highlighted in yellow. The highlighted text includes: 'Russia Sold \$6Bln Worth of Arms in 2005', 'Russia sold \$6 billion worth of arms', 'increase of \$500 million', 'military products', 'Russian Defense Minister Sergei Ivanov', 'President Vladimir Putin', '\$22 billion', 'orders', 'arms industry', 'Russia's plans', 'expand', 'sector', 'exports', 'products', 'Russian arms exports increased', 'Venezuela', 'Middle Eastern', 'Southeast Asian markets', 'Russia courted controversy when it said it would sell advanced surface-to-air missiles to Iran', and 'Washington and Israel'.

Architektura systému

- Segmentace a preprocesor
- Filtr
- Parser, Kombinace fragmentů
- Sémantická interpretace, odstranění víceznačností
- Generování šablon

Základní extrakce

- místa
 - oblasti, státy, města, ...
- osoby
 - stát, instituce, organizace, firmy, ...
- čísla
 - peníze, datum, telefonní čísla
- funkce

Komplexní extrakce

- obecně: entita – relace – entita
- instituce (firma) – funkce – osoba
- firma – obchod – zboží
- firma – obchod – zboží – firma
- atd.

stát – funkce – osoba

The screenshot displays a text extraction tool interface. At the top, the 'Input file content' section shows a paragraph of text: 'In 2005 Russia sold \$6 billion worth of arms, which is an increase of \$500 million on the previous year. Along with oil and gas, military products remain one of country's main exports. Russian Defense Minister Sergei Ivanov was quoted on Thursday, Jan. 19, as telling President Vladimir Putin: "The general volume of orders comes to \$22 billion at the moment, which means that our industry has serious orders for the next few years." The arms industry is key to Russia's plans to expand and diversify its economy. It is a rare sector of the economy where it can compete with Western nations in terms of exports of finished products. MosNews has reported in November 2005, that Russian arms exports increased 15-fold over the last three years. Among Russia's clients are Venezuela and Middle Eastern states. In recent years the country has been increasingly looking to Southeast Asian markets for further expansion. In December Russia courted controversy when it said it would sell advanced surface-to-air missiles to Iran, in a move harshly criticized by Washington and Israel.'

The 'Extraction result' section shows a tree structure of extracted entities. The root is 'securMilitary'. Under it, several categories are listed, including '/functionComplex/Russian defense Minister' and '/functionComplex/President|president Vladim'. Red arrows point from these categories to the corresponding text in the 'Extraction result' window on the right. The text in the window is highlighted in blue, and the names 'Russian Defense Minister Sergei Ivanov' and 'President Vladimir Putin' are highlighted in yellow. The text in the window is: 'Russia Sold \$6Bln Worth of Arms in 2005 In 2005 Russia sold \$6 billion worth of arms, which is an increase of \$500 million on the previous year. Along with oil and gas, military products remain one of country's main exports. Russian Defense Minister Sergei Ivanov was quoted on Thursday, Jan. 19, as telling President Vladimir Putin: "The general volume of orders comes to \$22 billion at the moment, which means that our industry has serious orders for the next few years." The arms industry is key to Russia's plans to expand and diversify its economy. It is a rare sector of the economy where it can compete with Western nations in terms of exports of finished products. MosNews has reported in November 2005, that Russian arms exports increased 15-fold over the last three years. Among Russia's clients are Venezuela and Middle Eastern states. In recent years the country has been increasingly looking to Southeast Asian markets for further expansion. In December Russia courted controversy when it said it would sell advanced surface-to-air missiles to Iran, in a move harshly criticized by Washington and Israel.'

stát – obchod – zbraně

The screenshot displays a software interface for text extraction. At the top, the 'Input file content' section shows the source text: 'Russia Sold \$6Bln Worth of Arms in 2005' followed by a paragraph about the increase in arms sales. An 'Extract' button is located to the right. Below this, the 'Extraction result' section features a tree view on the left and a text preview on the right. The tree view lists various semantic categories such as location, business operations, and specific entities like 'Russian Defense Minister Sergei Ivanov' and 'President Vladimir Putin'. The text preview on the right shows the original text with several phrases highlighted in yellow, corresponding to the categories in the tree view. Red arrows point from the tree view to the highlighted text in the preview.

Input file content

Russia Sold \$6Bln Worth of Arms in 2005

In 2005 Russia sold \$6 billion worth of arms, which is an increase of \$500 million on the previous year. Along with oil and gas, military products remain one of country's main exports.

Extract

Extraction result

- securMilitary
 - /locationArms/Russia sell \$6Bln Worth of arms
 - /locationName/Russia/Russia
 - /businessOper/sell
 - /dollarNum/\$6Bln
 - /weaponsN/Worth of arm
 - /locationArms/Russia sell \$6 billion worth of arms
 - /locationName/Russia/Russia
 - /businessOper/sell
 - /dollarNum/\$6
 - /ComplexWordNumber/billion
 - /weaponsN/worth of arm
 - /business/increase
 - /dollarComplex/\$500 million
 - /militaryBusiness/military product
 - /business/export
 - /functionComplex/Russian defense Minister
 - /functionComplex/President|president Vladimir Putin
 - /business/order
 - /dollarComplex/\$22 billion

Russia Sold \$6Bln Worth of Arms in 2005 In 2005 Russia sold \$6 billion worth of arms, which is an increase of \$500 million on the previous year. Along with oil and gas, military products remain one of country's main exports. Russian Defense Minister Sergei Ivanov was quoted on Thursday, Jan. 19, as telling President Vladimir Putin: "The general volume of orders comes to \$22 billion at the moment, which means that our industry has serious orders for the next few years." The arms industry is key to Russia's plans to expand and diversify its economy. It is a rare sector of the economy where it can compete with Western nations in terms of exports of finished products. MosNews has reported in November 2005, that Russian arms exports increased 15-fold over the last three years. Among Russia's clients are Venezuela and Middle Eastern states. In recent years the country has been increasingly looking to Southeast Asian markets for further expansion. In December Russia courted controversy when it said it would sell advanced surface-to-air missiles to Iran, in a move harshly criticized by Washington and Israel.

stát – obchod – zbraně – stát

The screenshot shows a software interface for text extraction. At the top, the 'Input file content' section contains two paragraphs of text. The first paragraph is: 'Among Russia's clients are Venezuela and Middle Eastern states. In recent years the country has been increasingly looking to Southeast Asian markets for futher expansion.' The second paragraph is: 'In December Russia courted controversy when it said it would sell advanced surface-to-air missiles to Iran, in a move'. An 'Extract' button is located to the right of the input text.

The 'Extraction result' section on the left displays a hierarchical tree of extracted terms. The terms include: '/business/sector', '/business/export', '/business/product', '/location/Russian', '/militaryBusiness/arm export', '/business/increase', '/location/Russia', '/business/client', '/location/Venezuela', '/location/middle eastern', '/location/southeast Asian', '/business/market', and '/rocketBusinessLocation/Russia court contro'. The last term is expanded to show sub-terms: '/locationName/Russia/Russia', '/businessOper/sell', '/nRockets/surface-to-air missile', and '/locationName/Iran/Iran'. A red arrow points to the expanded term.

The main text area on the right displays the extracted text in blue. A red arrow points to the highlighted sentence: 'Russia courted controversy when it said it would sell advanced surface-to-air missiles to Iran, in a move harshly criticized by Washington and Israel.'

Testování extrakcí

- Jazyky
 - Angličtina
 - Čeština
- Tematické oblasti
 - Bezpečnost
 - Competitive Intelligence

Závěry

- Flektivní jazyky obtížněji zpracovatelné nežli jazyky izolační
- Lingvistiku je třeba korigovat
- Odhadnout kdy ukončit rozvoj
 - dalším rozvojem – zhoršení výsledků
- Nasadit skutečně na extrakce
 - ne na kategorizaci, shlukování, atd.
kde je fulltext lepší