USING THE WEB TO MONITOR A CUSTOMIZED UNIFIED FINANCIAL PORTFOLIO

Camilo Restrepo-Arango
Arturo Henao-Chaparro
Claudia Jiménez-Guarín

Department of Systems and Computing Engineering
Universidad de los Andes - Bogotá, Colombia

Ninth International Workshop on Web Information Systems Modeling (WISM 2012)
31st International Conference on Conceptual Modeling (ER 2012)
October 15th – 18th, 2012, Florence, Italy
Agenda

- Motivation
- The Problem
- The Context
- Customized Unified Financial Portfolio
  - Web information integration features
  - Data model for customized queries
- The Proposed Solution
- Test and Results
- Conclusions and Future Work
Motivation

- Discovering of relevant and pertinent information on the Web
  - Flooding, discrimination, discovering, gathering
- Domain-specific searching and information integration
  - Financial personalized information
- Financial services and information for several audiences
  - Google Finance
  - Yahoo! Finance
  - Bloomberg
  - ...
- Usual general market trends information
  - Dynamic, factual
  - Usually stocks, news, RSS feeds
  - Ontologies, news analysis
  - Not social and crowdsourcing
- Lack of meaningful information integration
The Problem

- Filtering, gathering, discovering
  - On time
  - When needed
  - What is needed

- Not only financial facts are relevant
  - Analysis publishing, blogs, social media
  - Web 2.0 collaborative knowledge generation

- Personalization
  - User portfolios
  - Pertinent investments
  - More than stocks

- Perceptions, trends
The Context

- Use of Web information retrieval approaches
- Content is the more important feature
  - Integrated to structured data
  - Content is really more than tags and titles
- Sources are heterogeneous, autonomous
- No suppositions can be made about
  - Structure
  - Actual content description
  - Language quality
- Multilanguage sources
- Big data and Web 2.0 techniques are useful and appropriate
Customized Unified Financial Portfolio (UFP)

- Mash-up integrating 242 + 4 financial information sources
- Content is the basis for retrieval
- User financial portfolios filtering are considered
- Integration of detailed financial information, social trends and news for each portfolio asset.
- Big Data approach considering sources: V³
- Scalability, flexibility
- Information freshness and relevance vs. performance
- Content gathering, indexing, storing
- Semantic analysis considering social media sources
Information Integration Features

- 3 Categories of information, in 2 languages
  - Stock exchange sources: BVC, NYSE
  - Newspapers from several countries
    - Portafolio – Colombia
    - The Economist – England
    - New York Times – USA
    - The Wall Street Journal – USA
    - Crawled news feeds
  - Social networks: Facebook, Twitter
UFP process the information in three stages

1. Configuration
   A. initial interesting data sources
   B. Financial model established by a domain expert

2. Source crawling, information classification and indexing

3. Monitor of user relevant information concerning his portfolio

Example: Skandia Multifund Portfolios
UFP Application Architecture

- **UFPCore**
- **Data Extractor**
  - Specialized crawlers and analyzers for each information source type and language
  - Sentiment analysis of social media sources
- **Information Retrieval Model**
  - Full text search and indexing – Syntactic heterogeneity
  - Enterprise and relevant documents matching
- **NoSQL column store repository**
Financial Data Model

BVC, NYSE:
- RowKey[enterprise_name]
  - CF[Prices]
    - Column[InitialPrice]: Value
    - Column[MaximumPrice]: Value
    - Column[MinimumPrice]: Value
    - Column[MeanPrice]: Value
    - Column[LastPrice]: Value
    - Column[AnualVariation]: Value
    - Column[DalyVariation]: Value

SOCIAL_MEDIA:
- RowKey[Comment]
  - CF[Info]
    - Column[Name]: Value
    - Column[enterprise]: Value

SENTIMENTS:
- RowKey[id]
  - CF[Info]
    - Column[Sentiment]: Value
    - Column[enterprise]: Value
    - Column[news]: Value
    - Column[Comment]: Value
Asset Semantic Table Definition Process

Financial Expert Definition

Enterprise Asset: Ecopetrol

<table>
<thead>
<tr>
<th>Petroleum</th>
<th>1. Define relevant keywords associated with the industry.</th>
<th>8. Assign weights to every keyword associated with the industry</th>
<th>100%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exploration</td>
<td>2. Define relevant keywords associated with the main services of the company.</td>
<td>7. Assign weights for every keyword associated with the main services of the company</td>
<td>20%</td>
</tr>
<tr>
<td>Refinery Oil</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oil</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coil</td>
<td>3. Define relevant keywords associated with the rival business.</td>
<td>6. Assign weights to every keyword associated with the rival business.</td>
<td>40%</td>
</tr>
<tr>
<td>Environment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Renewable energy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EC</td>
<td>4. Define relevant keywords associated with the stock exchange.</td>
<td>5. Assign weights to every keyword associated with the stock exchange</td>
<td>70%</td>
</tr>
<tr>
<td>BVC NYSE</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Industry: 20% - Main Services: 30% Rival Business: 20% Stock: 30%
# Data Model for Customized Queries

## Ecopetrol

<table>
<thead>
<tr>
<th>Topic</th>
<th>Weight</th>
<th>Keywords</th>
<th>Keywords relative weight</th>
<th>Keywords absolute weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Industry (\text{ind})</td>
<td>(w_{\text{ind}}) 20%</td>
<td>(kw_{\text{ind}1}), (kw_{\text{ind}2}), (kw_{\text{ind}n}) Petroleum</td>
<td>(rkw_{\text{ind}1}), (rkw_{\text{ind}2}), (rkw_{\text{ind}n}) 100%</td>
<td>(tkw_{\text{ind}1}), (tkw_{\text{ind}2}), (tkw_{\text{ind}n})</td>
</tr>
<tr>
<td>Main Services (\text{ms})</td>
<td>(w_{\text{ms}}) 30%</td>
<td>(kw_{\text{ms}1}), (kw_{\text{ms}2}), (kw_{\text{ms}n})</td>
<td>(rkw_{\text{ms}1}), (rkw_{\text{ms}2}), (rkw_{\text{ms}n}) 100%</td>
<td>(tkw_{\text{ms}1}), (tkw_{\text{ms}2}), (tkw_{\text{ms}n})</td>
</tr>
<tr>
<td>Rival Business (\text{rb})</td>
<td>(w_{\text{rb}}) 20%</td>
<td>(kw_{\text{rb}1}), (kw_{\text{rb}2}), (kw_{\text{rb}n})</td>
<td>(rkw_{\text{ms}1}), (rkw_{\text{ms}2}), (rkw_{\text{ms}n}) 100%</td>
<td>(tkw_{\text{ms}1}), (tkw_{\text{ms}2}), (tkw_{\text{ms}n})</td>
</tr>
<tr>
<td>Stock Exchange (\text{se})</td>
<td>(w_{\text{se}}) 30%</td>
<td>(kw_{\text{se}1}), (kw_{\text{se}2}), (kw_{\text{se}n})</td>
<td>(rkw_{\text{ms}1}), (rkw_{\text{ms}2}), (rkw_{\text{ms}n}) 100%</td>
<td>(tkw_{\text{ms}1}), (tkw_{\text{ms}2}), (tkw_{\text{ms}n})</td>
</tr>
</tbody>
</table>
Document score in customized queries

\[ \text{topic} = [w_{\text{topic}}, \{k_{\text{topic}}w, r_{\text{topic}}k, t_{\text{topic}}w\}], \{k_{\text{topic}}w, r_{\text{topic}}k, t_{\text{topic}}w\} \neq \emptyset \] (1)

\[ \sum_{i \in \text{topics}} w_i = 1 \] (2)

\[ \sum_{i \in \text{KW}_t} r_{kw_i}t = 1, \forall t \in \text{topics} \] (3)

\[ tk_{kw_i}t = w_t \times r_{kw_i}t, \forall t \in \text{topics}, \forall i \in \text{KW}_t \] (4)

- When UFP retrieves the information about a specific asset, the semantic table is used to search not only the asset name, but also the associated keywords.

- Weights are used to calculate the final document score given an asset.

- Equations 2, 3, 4 define the relative importance of each topic keyword as well as the importance of an asset topic.
Document score in customized queries . . .

\[
UFPscore_{(q,d)} = C_{(q,d)} \cdot \text{norm}_{(q)} \cdot \text{sumt}_{(q,d)}
\]

\[
\text{norm}_q = \frac{1}{\sqrt{\sum_{t \in q} (idf(t) \cdot tkw(t))^2}}
\]

\[
\text{sumt}_{(q,d)} = \sum_{t \in q} (tf(t,d) \cdot idf(t)^2 \cdot tkw(t))
\]

- The Vector-Space Model \(C_{(q,d)}\): how many keywords are found in the document
- \(\text{norm}_{(q)}\): normalizing factor making comparable scores between queries, using the inverse document frequency sum for all the terms in order to convert the final score to a normal form
- \(\text{sumt}_{(q,d)}\): contribution of each keyword to the query score. It is based in the frequency of a keyword in the document, the inverse document frequency where rarer keywords get higher scores and the total weight of the keyword (eq. 7).
The Proposed Solution – Web Interface

Una patadita de la buena suerte a Millionarios

Bajo de dinero Lea una columna de opinión de Juan Carlos Ortiz, un joven pero experimentado ex corredor de bolsa, de quién se dice es el hombre fuerte detrás de los grupos financieros Proyector Valores e Interbolsa, retrato de la extinta Bolsa de Bogotá, y quien casi que de una manera inocente lanzó ayer la siguiente pregunta: ”Y qué pasa si Millonarios va a la bolsa? leen más

Banco do Brasil atemoriza esto aqo en Colombia

Este interesado en prestar sus servicios a las compañías brasileñas que están en la presencia en el país. La llegada de Banco do Brasil a Colombia será una realidad este año, a través de la compra de una entidad del sistema financiero local o por medio de la apertura de una oficina de representación, que prestará sus servicios a las compañías brasileñas que están en el país. Así lo anunció el presidente de esta entidad, Carlos Nascimento, quien llegó a Bogotá a anunciar una alianza con la Sociedad Administradora de Inversión del Grupo Interbolsa. leer más

InterBolsa expande su portafolio y seré una fiduciaria

La administradora de inversiones del Grupo tramita la conversión a fiduciaria. Espera duplicar sus ingresos. Las inversiones en activos diferentes a lo que tradicionalmente ofrece el mercado de capitales será una de las apuestas del Grupo InterBolsa para los próximos meses. La construcción de un centro comercial en Santa Marta, la originación de hipotecas, la creación de un fondo de pensiones voluntarias y hasta la microfinanzas, son algunas de las iniciativas que ya están en trámite o que tienen entre los planes de mediano plazo. leer más
Personalization

- The Web is crawled for all of the defined assets
  - All interesting assets can be described and included
  - Interesting information sources can be defined as crawling seeds
  - Related interesting information sources are automatically discovered

- User defines his portfolio composition
  - Assets and amount of money in each personal investment alternative
  - The system calculates relative assets weights

- News and results are filtered and displayed considering personal investments
  - Filtering of personal information at query time
Tests and Results

- Fully implemented prototype
  - Debian 6
  - Java, Glassfish
  - Apache Hbase
  - Synesketch

- Data Extractor recollected data over 52 days
  - 38 working days for the stocks exchange data
  - 29919 financial news
  - 3152 elements from Social Media (mainly Twitter)

- Displayed information considers relationships through business domain for the actual user on query

- Information retrieval measures are taken over a sample of a hundred of documents, classified manually
Example for news content over 4 colombian enterprises:

<table>
<thead>
<tr>
<th>Company</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Odinsa</td>
<td>70%</td>
<td>84%</td>
</tr>
<tr>
<td>Interbolsa</td>
<td>28.5%</td>
<td>44%</td>
</tr>
<tr>
<td>Davivienda</td>
<td>75%</td>
<td>76%</td>
</tr>
<tr>
<td>Ecopetrol</td>
<td>69.9%</td>
<td>88%</td>
</tr>
</tbody>
</table>

Comparable with the Hermes Framework and YourNews related works

Good recall and precision for enterprises centered in one industry: Odinsa, Davivienda, Ecopetrol

Poor precision for enterprises involved in several topics

Synesketch sentiment analysis is not customizable, results are not domain-specific

LingPipe is a better option for this component, even if training is needed
Conclusions

- Successful integration of non-structured heterogeneous and domain-specific content
- More than 240 information sources
  - Multilanguage, international and regional coverage
  - Social media, specialized sources, news integration
  - Web and structured public data
- Specific domain model based on vector-space model
- Customized and configurable both in domain and user dimensions
Conclusions...

- Classification and information retrieval techniques in order to deliver appropriate content
- Sentiment analysis and perception of financial information is included
- NoSQL technology for scalability and flexibility
- The proposed architecture can be used in other domain contexts
- The semantic table can be defined for other domains
Questions?

Contact:
Camilo Restrepo-Arango
Arturo Henao-Chaparro
Claudia Jiménez-Guarín

Email: c.restrepo235@uniandes.edu.co
a.henao59@uniandes.edu.co
cjimenez@uniandes.edu.co

http://comit.uniandes.edu.co
Blog: http://wordpress1.virtual.uniandes.edu.co