Contextualizing data warehouses with documents

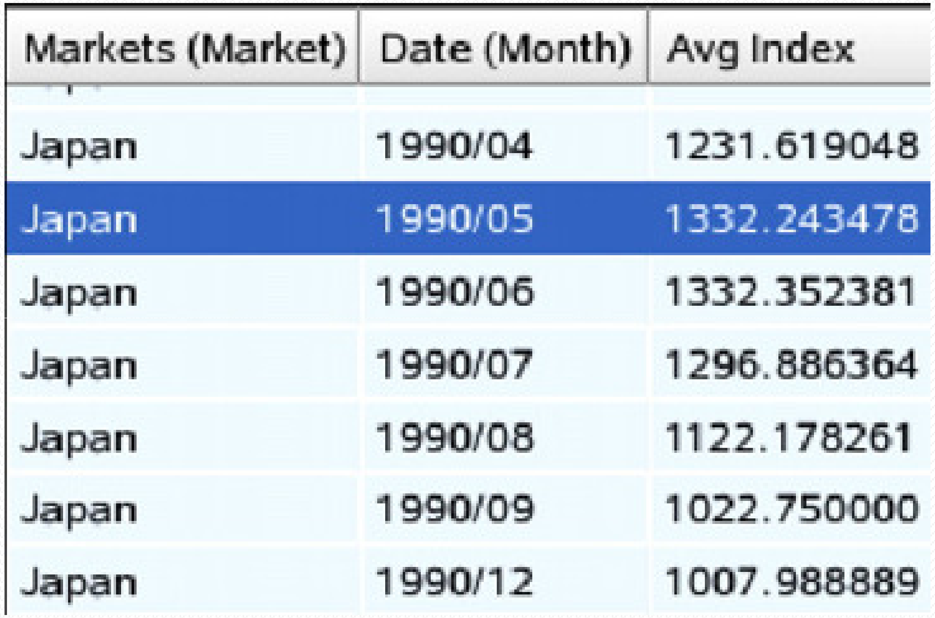
Written by:
Juan Manuel Pérez-Martínez
Rafael Berlanga-Llavori
María José Aramburu-Cabo
Torben Bach Pedersen
Published:
Available online from 7 February 2007
Journal by Elsevier from April 2008
Presenter:
Peter G. Poulsen

Presentation Overview

- Motivating Example
- Architecture
- Components
- R-cube
- Algebra
- Prototype
- Conclusion
- Related Works
- Evaluation

Motivation

- Stock Index dropped
- Why? Missing Context

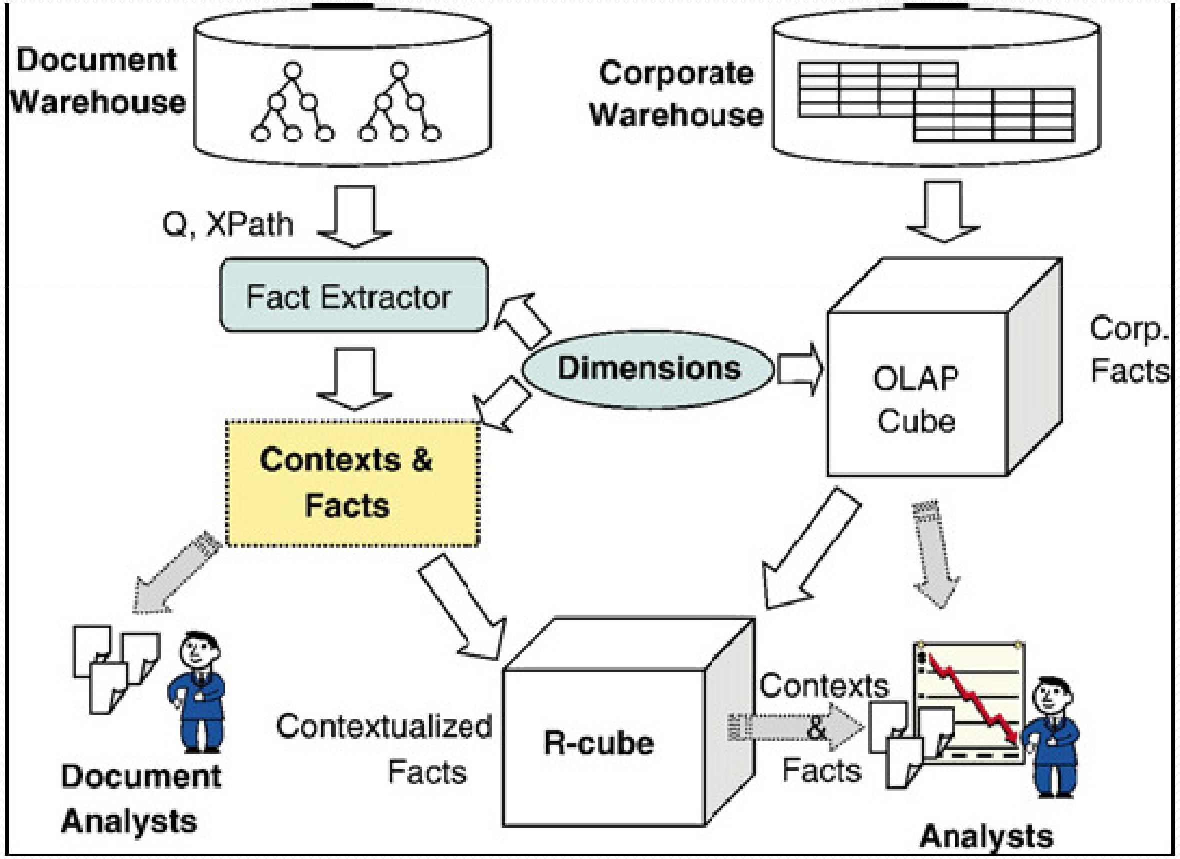


Markets (Market)	Date (Month)	Avg Index
Japan	1990/04	1231.619048
Japan	1990/05	1332.243478
Japan	1990/06	1332.352381
Japan	1990/07	1296.886364
Japan	1990/08	1122.178261
Japan	1990/09	1022.750000
Japan	1990/12	1007.988889

Motivation

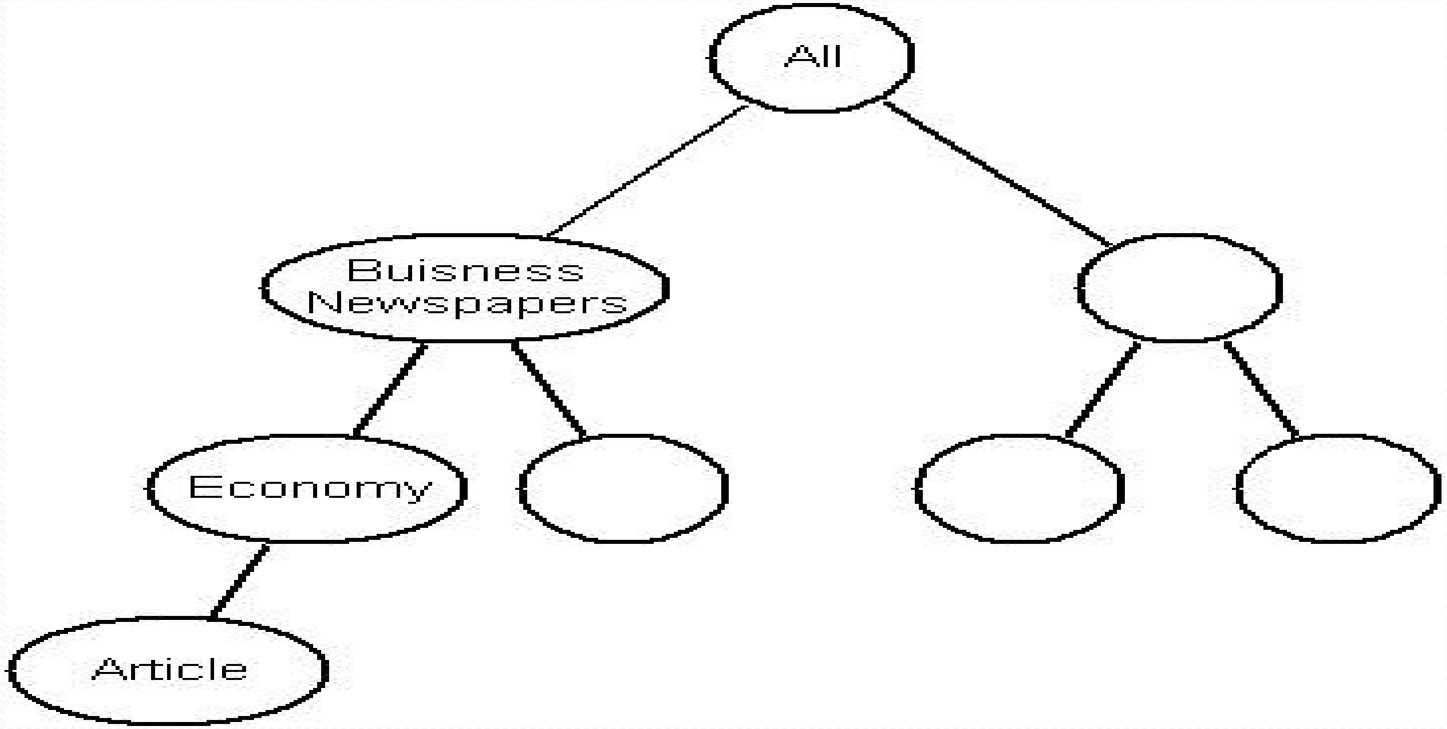
- Plant engineering companies fell sharply as their activities in Iraq and Kuwait have been frozen by Japan’s economic sanctions against Iraq. Chiyoda lost 150 to 1660.
- Similar behavior if same context appears
- Link facts and documents

Architecture



Document Warehouse

- XML-tree Structure of Documents



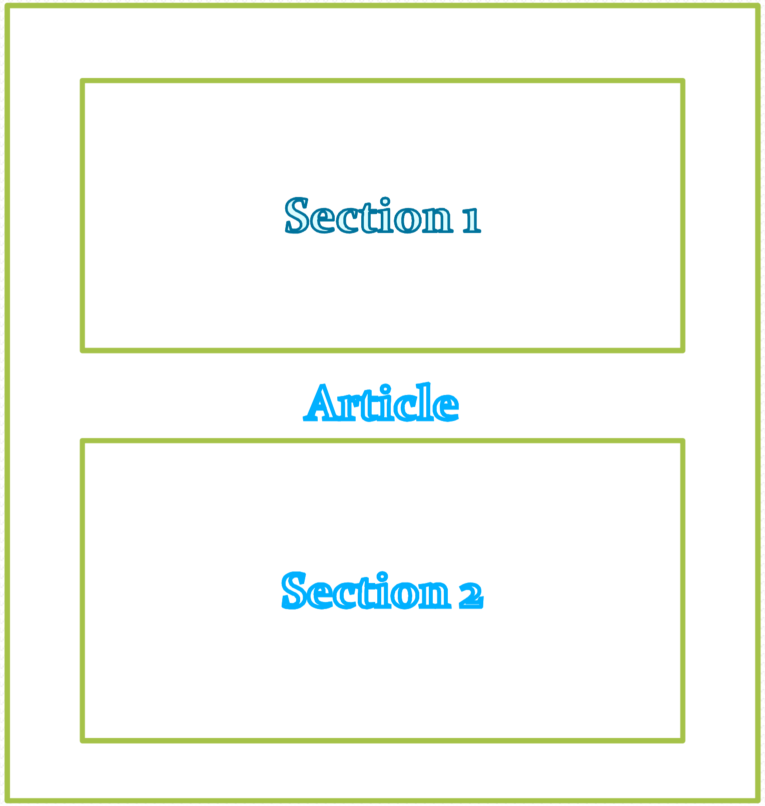
Document Warehouse

- Input on the Form: (XPath, Q)
  - XPath - restriction on the XML tree
  - Q - sequence of keywords

- Output: RQ (set of document nodes)
  - They are selected by XPath
  - They contain m(user specified) keywords (from Q)
  - They are more relevant than their subtrees

Document Warehouse

- Article
  - 100 words
  - 10 keywords
- Section 1
  - 50 words
  - 2 keywords
- Section 2
  - 50 words
  - 8 keywords

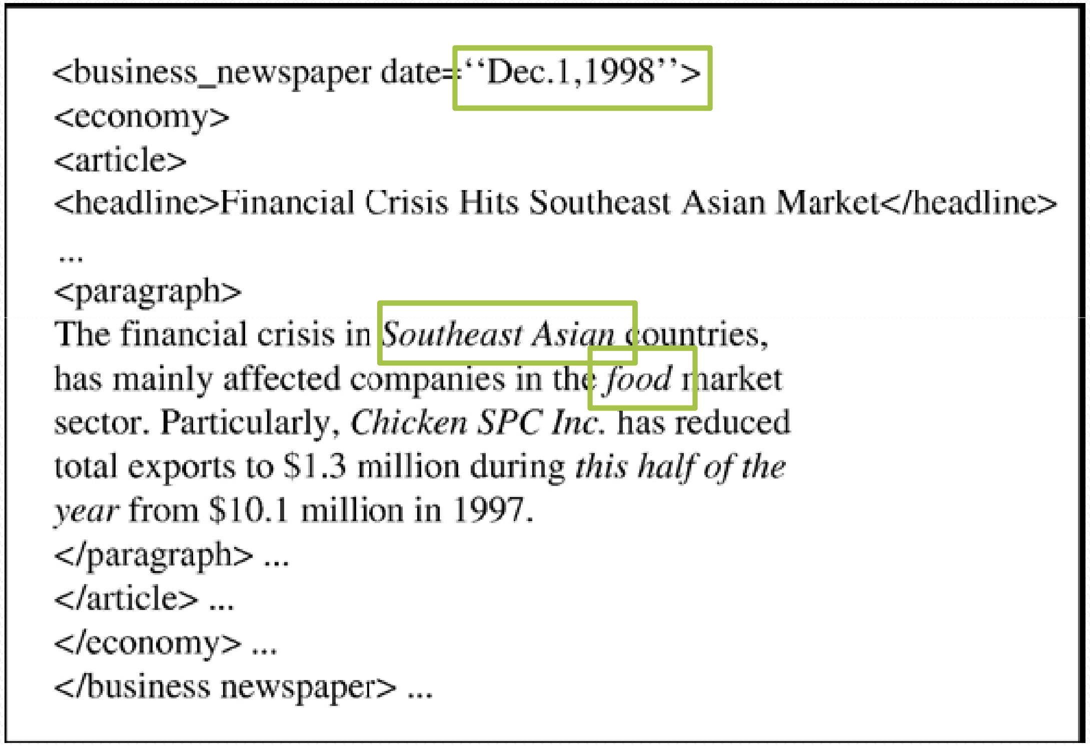


Fact Extractor

- Input RQ from the document warehouse
- Analyses the dimensions in the data warehouse
- Builds all facts in the documents based on the schema of the corporate warehouse

Fact Extractor

- 3 Dimensions
  - Time
  - Product
    - Food
    - Healthcare
  - Customers
    - Countries
    - Regions



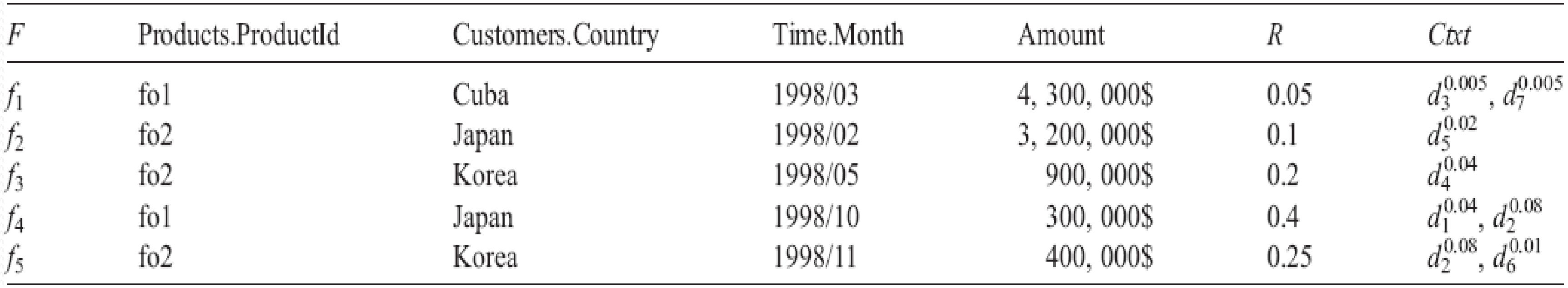
```
<business_newspaper date=''Dec.1,1998''>
<economy>
<article>
<headline>Financial Crisis Hits Southeast Asian Market</headline>
...
<paragraph>
The financial crisis in Southeast Asian countries,
has mainly affected companies in the food market
sector. Particularly, Chicken SPC Inc. has reduced
total exports to $1.3 million during this half of the
year from $10.1 million in 1997.
</paragraph> ...
</article> ...
</economy> ...
</business newspaper> ...
```

R-cube

- Input (XPath, Q, MDX)
- 5 steps to build the R-cube
  - 1. XPath and Q are evaluated on the document warehouse. Giving RQ.
  - 2. Facts are extracted from RQ along with frequencies.
  - 3. MDX is evaluated on the corporate warehouse.
  - 4. Documents are assigned to the facts, where the dimensions can be rolled up or drilled down to the facts described by the documents.
  - 5. Relevance of each fact is calculated.

R-cube

- Q="financial, crisis",
- XPath="/business_newspaper/economy/article//"
- MDX=(Products.[food], Customers.Country, Time.[1998].Month, SUM(Measures.Amount)>0)
- Only food products, customer countries, months of 1998 and the measures which sum is above 0.



F	Products.ProductId	Customers.Country	Time.Month	Amount	R	$Ctxt$
f_1	fo1	Cuba	1998/03	4, 300, 000$	0.05	$d_3^{0.005}, d_7^{0.005}$
f_2	fo2	Japan	1998/02	3, 200, 000$	0.1	$d_5^{0.02}$
f_3	fo2	Korea	1998/05	900, 000$	0.2	$d_4^{0.04}$
f_4	fo1	Japan	1998/10	300, 000$	0.4	$d_1^{0.04}, d_2^{0.08}$
f_5	fo2	Korea	1998/11	400, 000$	0.25	$d_2^{0.08}, d_6^{0.01}$

R-cube

- The relevance R of a fact f is:

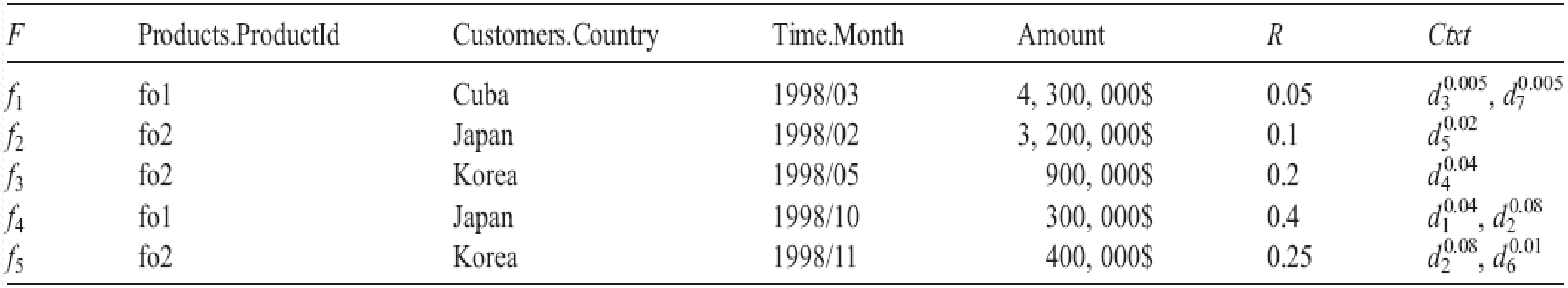
$$P(f \mid RQ) = \frac{\sum_{d \in RQ} P(f \mid d) P(Q \mid d)}{\sum_{d \in RQ} P(Q \mid d)}$$

$$P(f \mid d) = \frac{FF(f, d)}{|d|_f}$$

R-cube

- Relevance Example:
- d3 – 100 facts, 4 is f1, d7 – 100 facts, 6 is f1

$$P(f_1 \mid RQ) = \frac{0.04 \cdot 0.005 + 0.06 \cdot 0.005}{0.005 + 0.005} = 0.05$$



F	Products.ProductId	Customers.Country	Time.Month	Amount	*R*	*Ctxt*
f_1	fo1	Cuba	1998/03	4, 300, 000$	0.05	$d_3^{0.005}, d_7^{0.005}$
f_2	fo2	Japan	1998/02	3, 200, 000$	0.1	$d_5^{0.02}$
f_3	fo2	Korea	1998/05	900, 000$	0.2	$d_4^{0.04}$
f_4	fo1	Japan	1998/10	300, 000$	0.4	$d_1^{0.04}, d_2^{0.08}$
f_5	fo2	Korea	1998/11	400, 000$	0.25	$d_2^{0.08}, d_6^{0.01}$

Defining the R-cube

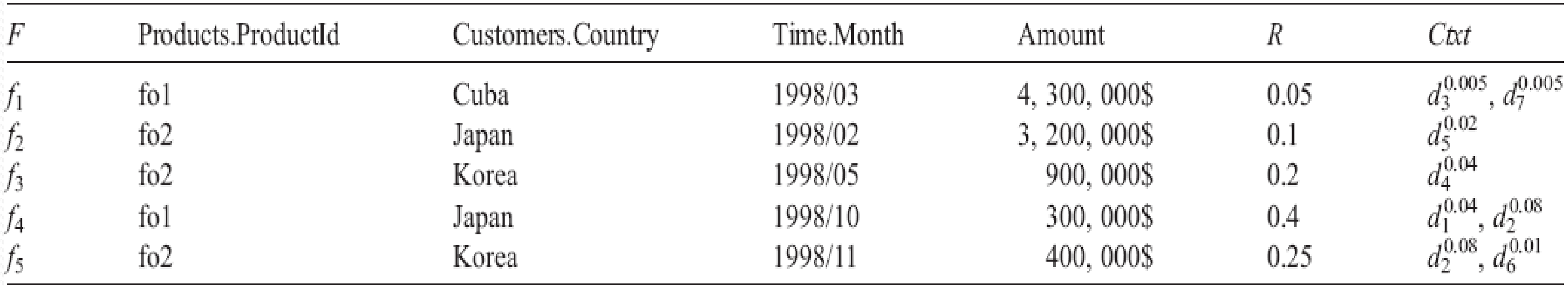
- Relevance Dimension
- Relevance-Fact Relation
- Context Dimension
- Context-Fact Relation
- R-cube definition

Relevance Dimension

- Real Number R– [0, 1]
- Relevance Degree – Very Relevant, Relevant etc
- Split the space [0, 1] into pieces each representing a relevance degree value
- Map R to relevance degree - $\frac{R}{\gamma}$

Relevance Dimension

- Example: $\gamma = MAX(R)$ and relevance degree is split into 5 values. Very irrelevant – [0, 0.25[, irrelevant – [0.25, 0.45[, neutral – [0.45, 0.55[, relevant – [0.55, 0.75[and very relevant – [0.75, 1].



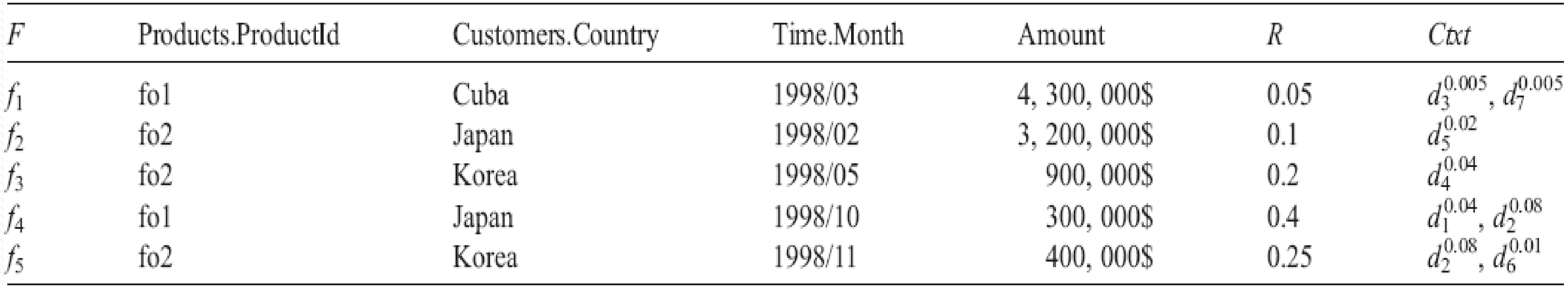
F	Products.ProductId	Customers.Country	Time.Month	Amount	R	$Ctxt$
f_1	fo1	Cuba	1998/03	4, 300, 000$	0.05	$d_3^{0.005}, d_7^{0.005}$
f_2	fo2	Japan	1998/02	3, 200, 000$	0.1	$d_5^{0.02}$
f_3	fo2	Korea	1998/05	900, 000$	0.2	$d_4^{0.04}$
f_4	fo1	Japan	1998/10	300, 000$	0.4	$d_1^{0.04}, d_2^{0.08}$
f_5	fo2	Korea	1998/11	400, 000$	0.25	$d_2^{0.08}, d_6^{0.01}$

- f1 – very irrelevant, f2 - irrelevant, f3 – relevant, etc

Relevance Fact-Dimension Relation

- FD={(f, R)}, f – fact and R – relevance
- FD={(f1, 0.05), …, (f5, 0.25)}

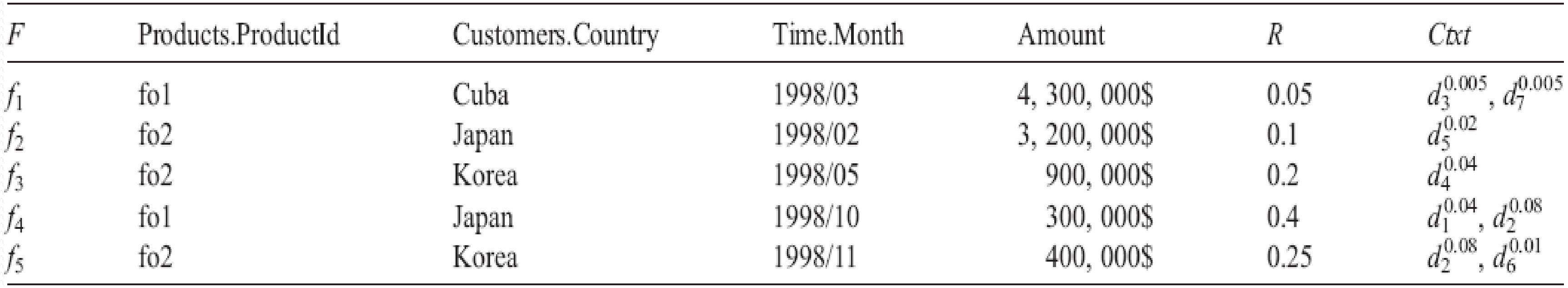
- $f \rightarrow_R^{\gamma} rd$, f – fact, rd – relevance degree and γ - global relevance measure
- $f_1 \rightarrow_R^{MAX(R)} very - irrelevant$



F	Products.ProductId	Customers.Country	Time.Month	Amount	R	$Ctxt$
f_1	fo1	Cuba	1998/03	4, 300, 000$	0.05	$d_3^{0.005}, d_7^{0.005}$
f_2	fo2	Japan	1998/02	3, 200, 000$	0.1	$d_5^{0.02}$
f_3	fo2	Korea	1998/05	900, 000$	0.2	$d_4^{0.04}$
f_4	fo1	Japan	1998/10	300, 000$	0.4	$d_1^{0.04}, d_2^{0.08}$
f_5	fo2	Korea	1998/11	400, 000$	0.25	$d_2^{0.08}, d_6^{0.01}$

Context Dimension

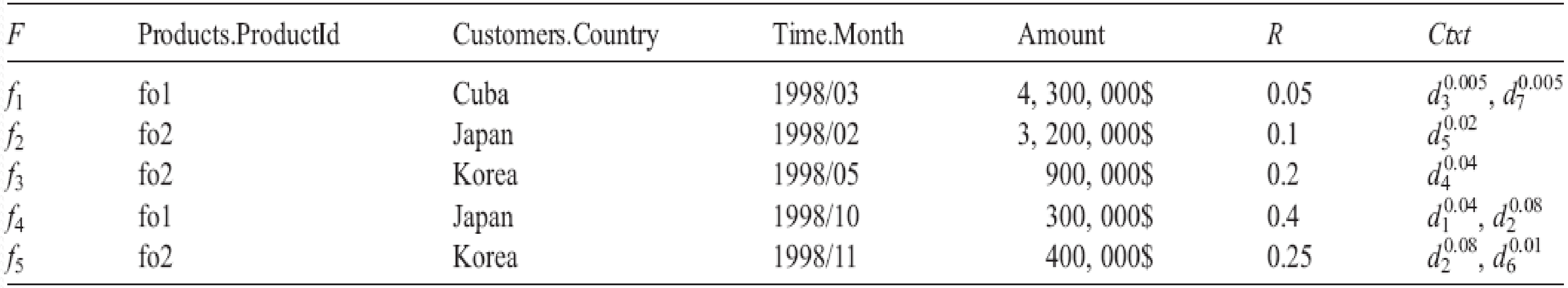
- Documents which describe the context
- Superscript is the relevance in relation to Q(the context)
- Example: $d_1^{0.04}$



F	Products.ProductId	Customers.Country	Time.Month	Amount	R	$Ctxt$
f_1	fo1	Cuba	1998/03	4, 300, 000$	0.05	$d_3^{0.005}, d_7^{0.005}$
f_2	fo2	Japan	1998/02	3, 200, 000$	0.1	$d_5^{0.02}$
f_3	fo2	Korea	1998/05	900, 000$	0.2	$d_4^{0.04}$
f_4	fo1	Japan	1998/10	300, 000$	0.4	$d_1^{0.04}, d_2^{0.08}$
f_5	fo2	Korea	1998/11	400, 000$	0.25	$d_2^{0.08}, d_6^{0.01}$

Context Fact-Dimension Relation

- $FC_{txt} = \{(f, d)\}$, f – fact and d is a document node
- $FC_{txt} = \{(f_1, d_3^{0.005}), (f_1, d_7^{0.005}), \ldots, (f_5, d_6^{0.01})\}$



F	Products.ProductId	Customers.Country	Time.Month	Amount	R	$Ctxt$
f_1	fo1	Cuba	1998/03	4, 300, 000$	0.05	$d_3^{0.005}, d_7^{0.005}$
f_2	fo2	Japan	1998/02	3, 200, 000$	0.1	$d_5^{0.02}$
f_3	fo2	Korea	1998/05	900, 000$	0.2	$d_4^{0.04}$
f_4	fo1	Japan	1998/10	300, 000$	0.4	$d_1^{0.04}, d_2^{0.08}$
f_5	fo2	Korea	1998/11	400, 000$	0.25	$d_2^{0.08}, d_6^{0.01}$

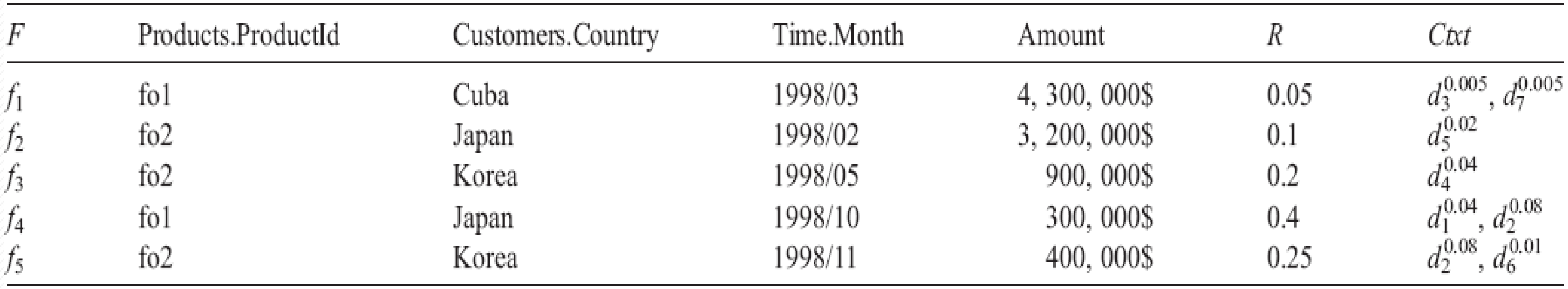
R-cube

- Four-tuple (F, D, FD, Q)
- F is the set of facts
- D is the set of dimensions, including relevance and context dimensions
- FD is the set of relations, including the relevance-fact and context-fact relations
- Q is IR condition

- Quality: Sum of document relevance to Q

R-cube Algebra

- Selection
  - Modify Relevance
  - Modify Quality
- Projection
  - Cannot remove Relevance or Context
- Aggregation
  - Sum Relevance
  - Union Context

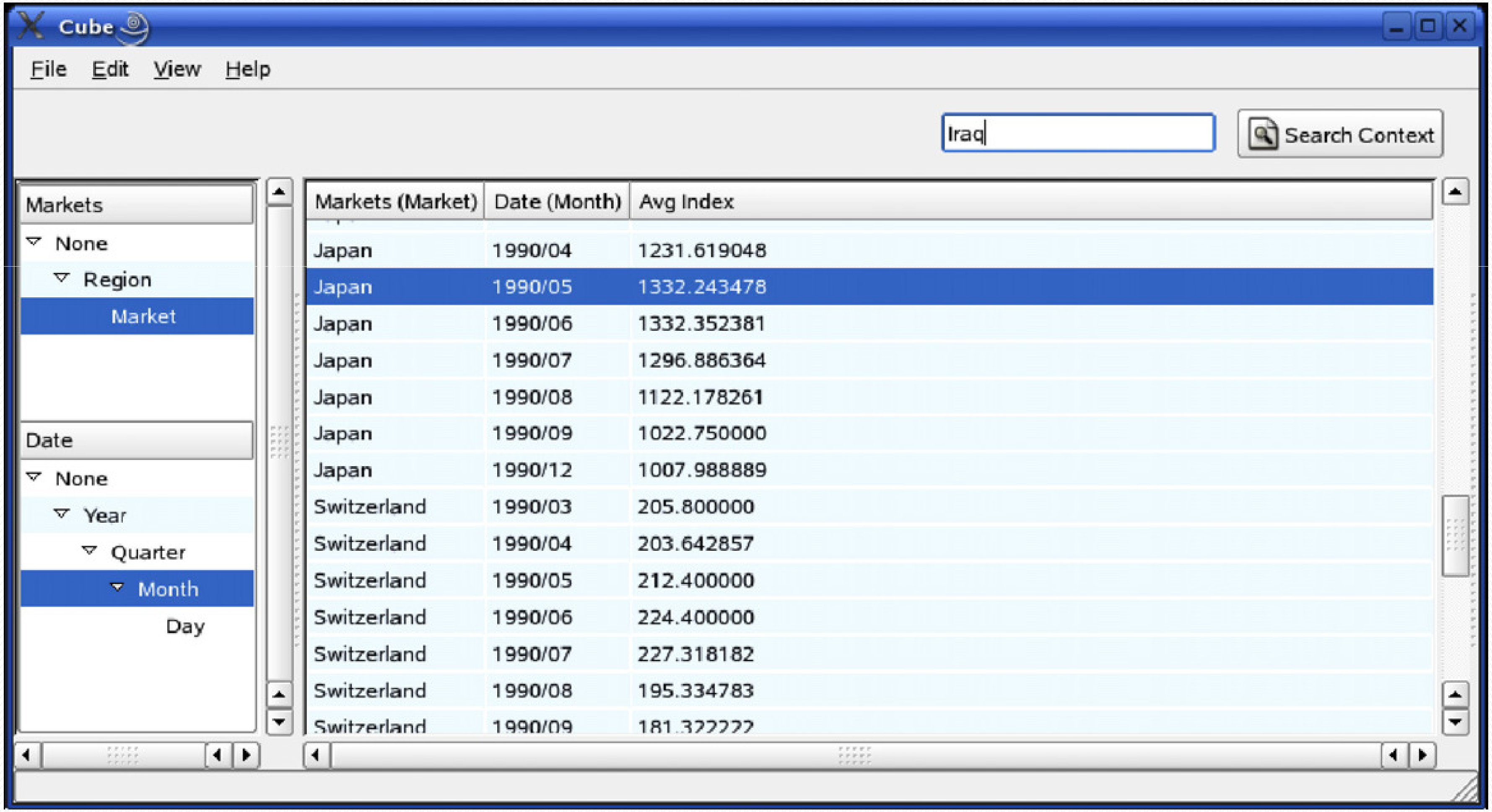


F	Products.ProductId	Customers.Country	Time.Month	Amount	R	$Ctxt$
f_1	fo1	Cuba	1998/03	4, 300, 000$	0.05	$d_3^{0.005}, d_7^{0.005}$
f_2	fo2	Japan	1998/02	3, 200, 000$	0.1	$d_5^{0.02}$
f_3	fo2	Korea	1998/05	900, 000$	0.2	$d_4^{0.04}$
f_4	fo1	Japan	1998/10	300, 000$	0.4	$d_1^{0.04}, d_2^{0.08}$
f_5	fo2	Korea	1998/11	400, 000$	0.25	$d_2^{0.08}, d_6^{0.01}$

Prototype

- 132 articles from 1990
- 1396 facts at lowest index categories
- 2 dimensions
  - Market
    - Market and Region
  - Date
    - Year, Quarter, Month and Day
- Context: Iraq

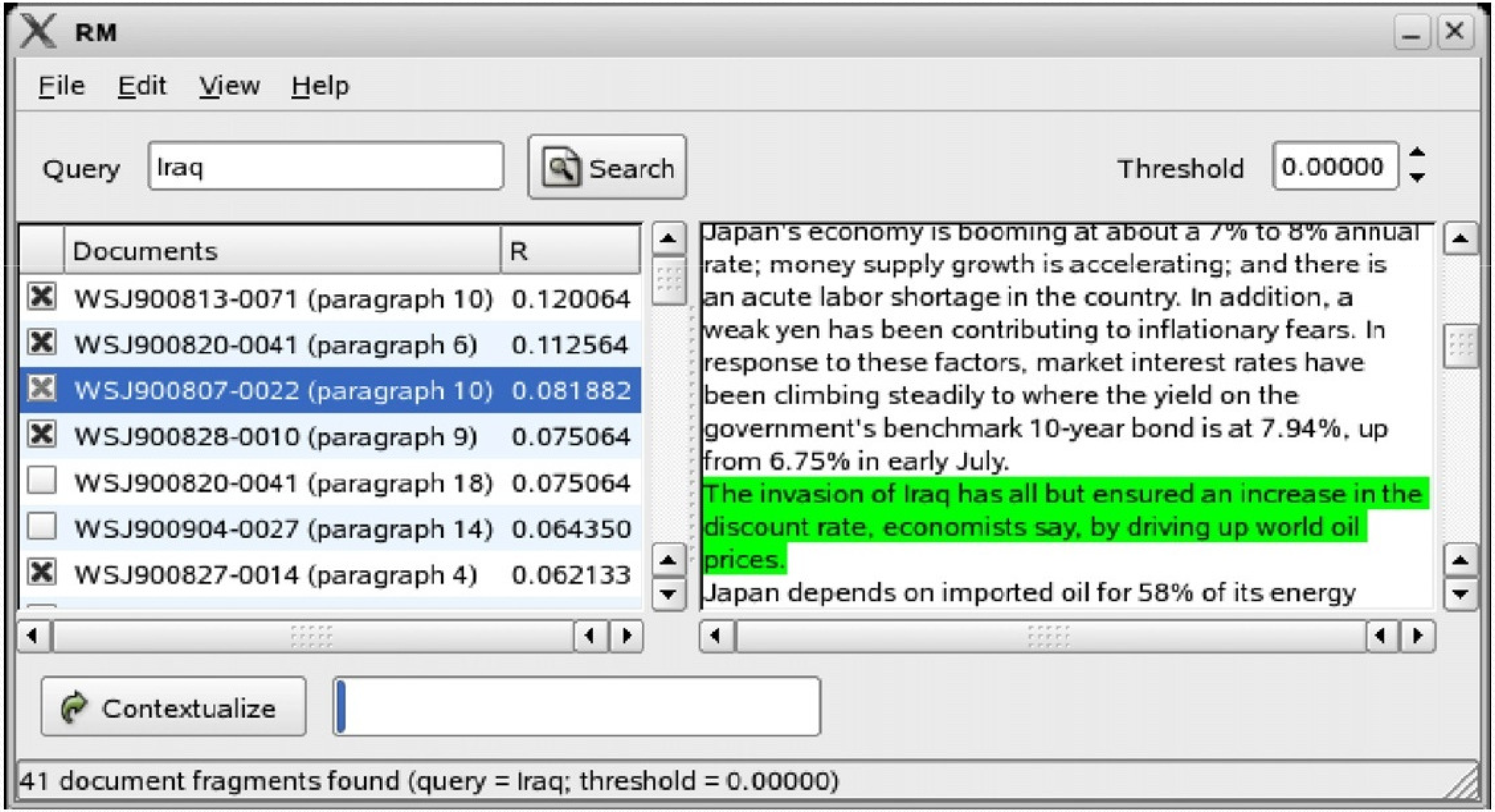
Prototype



Iraq

Markets (Market)	Date (Month)	Avg Index
Japan	1990/04	1231.619048
Japan	1990/05	1332.243478
Japan	1990/06	1332.352381
Japan	1990/07	1296.886364
Japan	1990/08	1122.178261
Japan	1990/09	1022.750000
Japan	1990/12	1007.988889
Switzerland	1990/03	205.800000
Switzerland	1990/04	203.642857
Switzerland	1990/05	212.400000
Switzerland	1990/06	224.400000
Switzerland	1990/07	227.318182
Switzerland	1990/08	195.334783
Switzerland	1990/09	181.322222

Prototype



Query: Iraq

Threshold: 0.00000

	Documents	R
☒	WSJ900813-0071 (paragraph 10)	0.120064
☒	WSJ900820-0041 (paragraph 6)	0.112564
☒	WSJ900807-0022 (paragraph 10)	0.081882
☒	WSJ900828-0010 (paragraph 9)	0.075064
□	WSJ900820-0041 (paragraph 18)	0.075064
□	WSJ900904-0027 (paragraph 14)	0.064350
☒	WSJ900827-0014 (paragraph 4)	0.062133

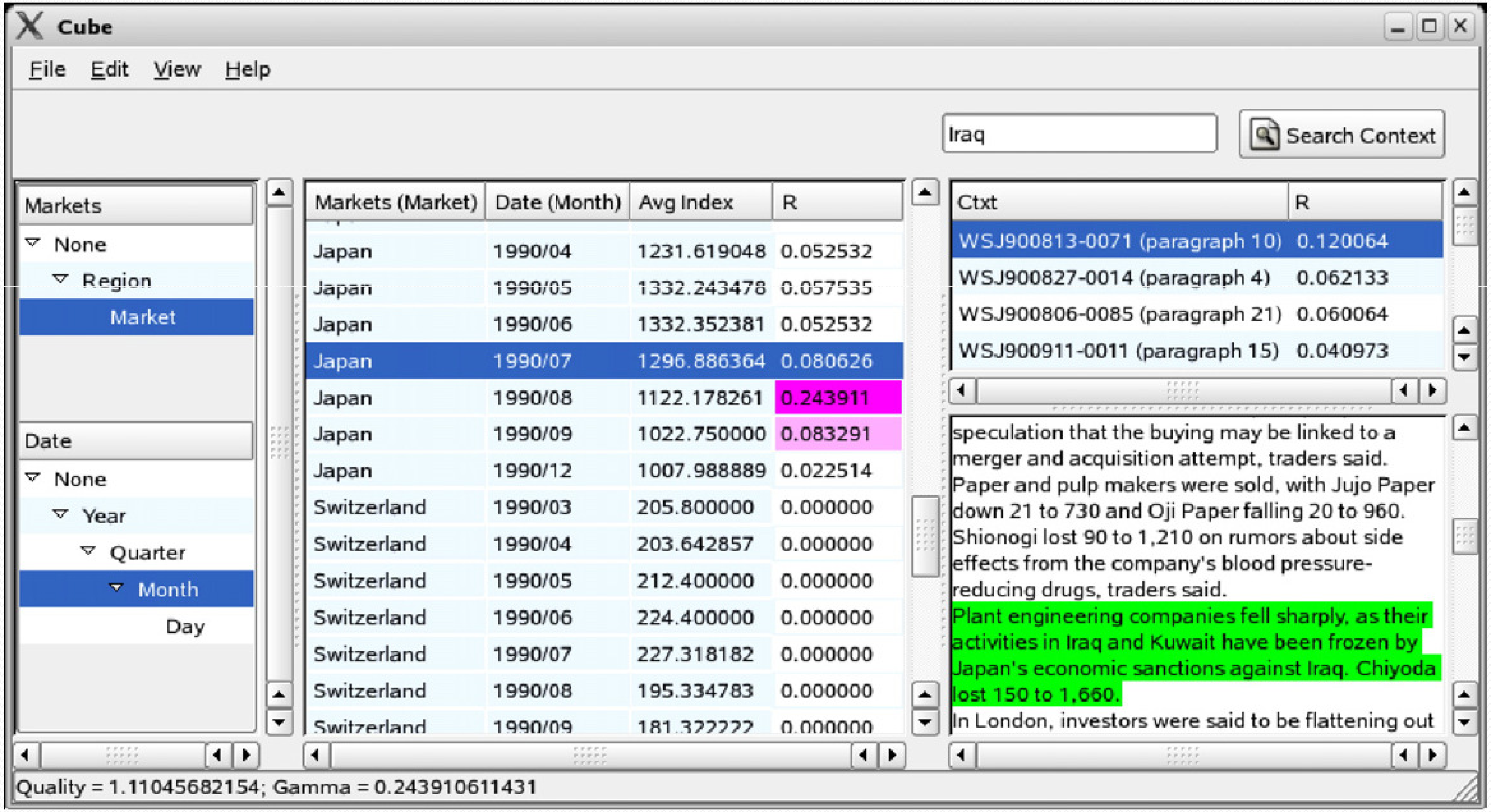
Japan's economy is booming at about a 7% to 8% annual rate; money supply growth is accelerating; and there is an acute labor shortage in the country. In addition, a weak yen has been contributing to inflationary fears. In response to these factors, market interest rates have been climbing steadily to where the yield on the government's benchmark 10-year bond is at 7.94%, up from 6.75% in early July.

The invasion of Iraq has all but ensured an increase in the discount rate, economists say, by driving up world oil prices.

Japan depends on imported oil for 58% of its energy

41 document fragments found (query = Iraq; threshold = 0.00000)

Prototype



Markets (Market)	Date (Month)	Avg Index	R
Japan	1990/04	1231.619048	0.052532
Japan	1990/05	1332.243478	0.057535
Japan	1990/06	1332.352381	0.052532
Japan	1990/07	1296.886364	0.080626
Japan	1990/08	1122.178261	0.243911
Japan	1990/09	1022.750000	0.083291
Japan	1990/12	1007.988889	0.022514
Switzerland	1990/03	205.800000	0.000000
Switzerland	1990/04	203.642857	0.000000
Switzerland	1990/05	212.400000	0.000000
Switzerland	1990/06	224.400000	0.000000
Switzerland	1990/07	227.318182	0.000000
Switzerland	1990/08	195.334783	0.000000
Switzerland	1990/09	181.322222	0.000000

Ctxt	R
WSJ900813-0071 (paragraph 10)	0.120064
WSJ900827-0014 (paragraph 4)	0.062133
WSJ900806-0085 (paragraph 21)	0.060064
WSJ900911-0011 (paragraph 15)	0.040973

speculation that the buying may be linked to a merger and acquisition attempt, traders said. Paper and pulp makers were sold, with Jujo Paper down 21 to 730 and Oji Paper falling 20 to 960. Shionogi lost 90 to 1,210 on rumors about side effects from the company's blood pressure-reducing drugs, traders said.

Plant engineering companies fell sharply, as their activities in Iraq and Kuwait have been frozen by Japan's economic sanctions against Iraq. Chiyoda lost 150 to 1,660.

In London, investors were said to be flattening out

Quality = 1.11045682154; Gamma = 0.243910611431

Conclusion

- Contextualized Warehouse
    - More detail, linking unstructured data in documents to structured data in the corporate warehouse
    - Architecture of the combined corporate and document warehouse
- Defined a R-cube with relevance and context dimensions
- Created a prototype to illustrate the use of the solution

Related Works

- Further development of earlier article on a model for text rich XML documents from the authors
- Applied relevance modeling

- Other approaches only deal with highly structured XML documents
- Nothing have been done with unstructured documents before

Relation to our Project

- Data Warehouse for the healthcare sector in Herning Kommune
- Data in the form of comments – 2 kinds
  - Comments to structured data
  - Stand alone comments
- Way to link structured and unstructured data

Evaluation

- Easily understandable paper
  - Good figures
  - Many examples
- Good flow

- Missing some performance measures
  - Longer query time? Compared to a normal data warehouse