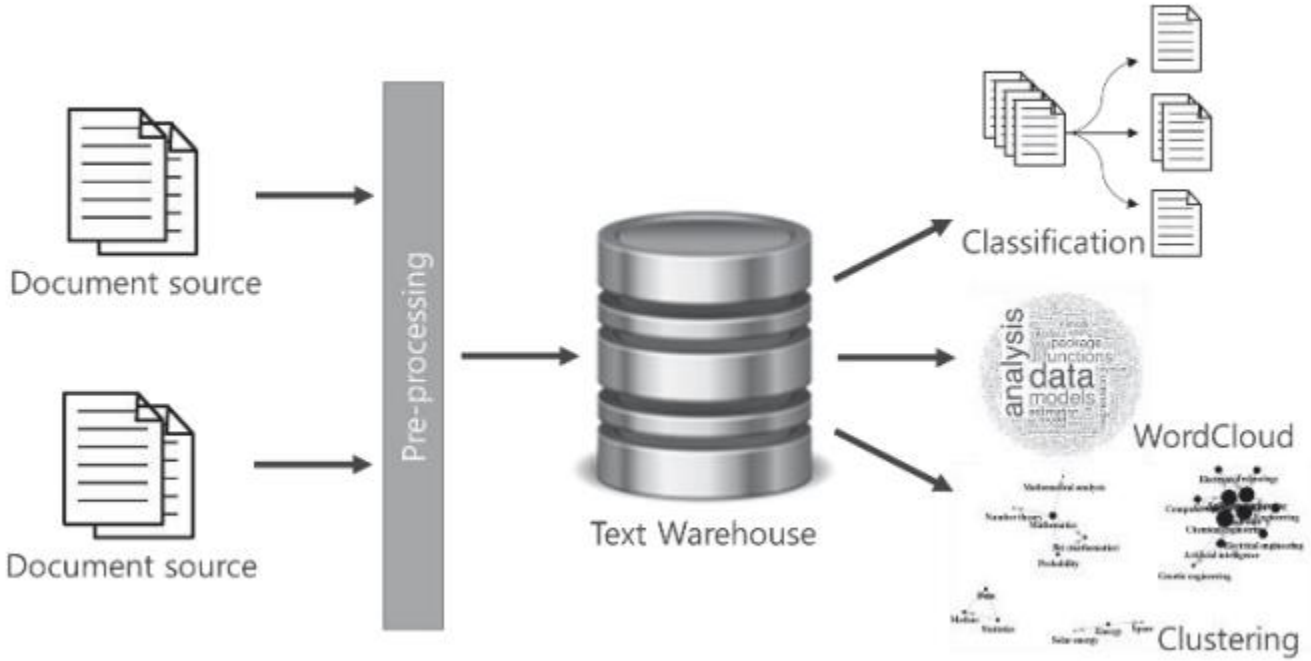
Text Warehousing

2017. 5

Introduction

- Big Text Data
  - The textual data integration in OLAP analysis and aggregating them to improve the decision-making is a challenge in Business Intelligence systems.
  - important to define new aggregation techniques appropriate to textual data.
- Proposed technique
  - IR techniques together with OLAP to better analyze textual data and extract their semantics.

Architecture



Text Warehouse 예

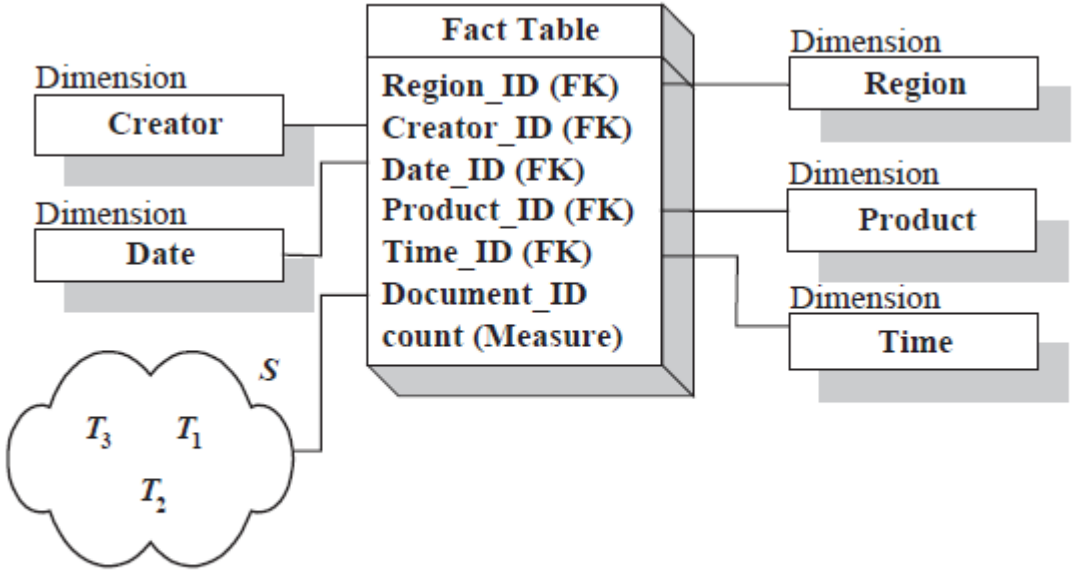


Fig. 9. An example star schema for complaint e-mail management.

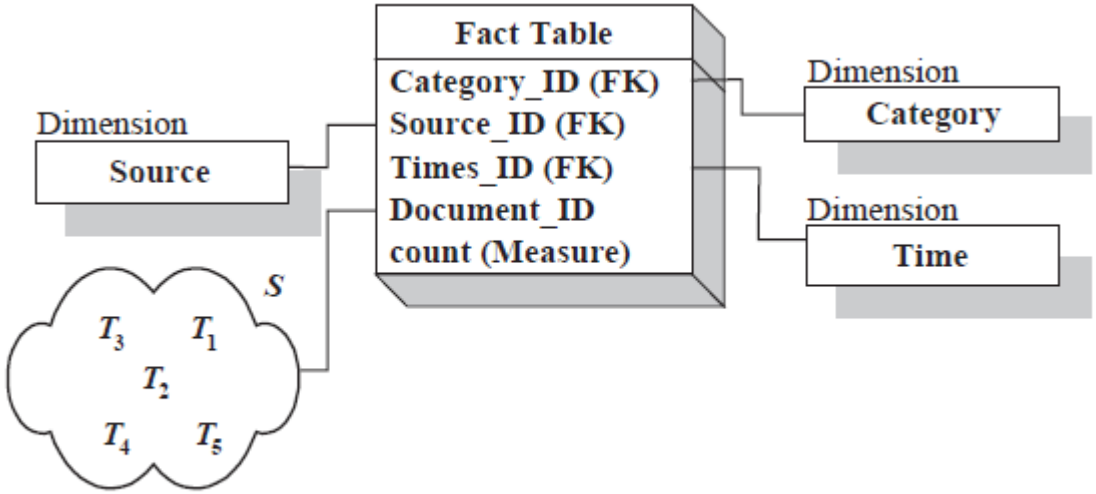
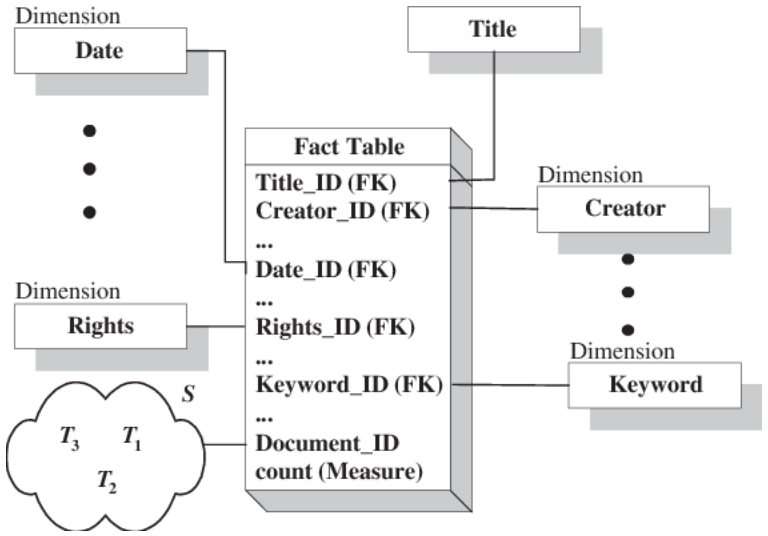
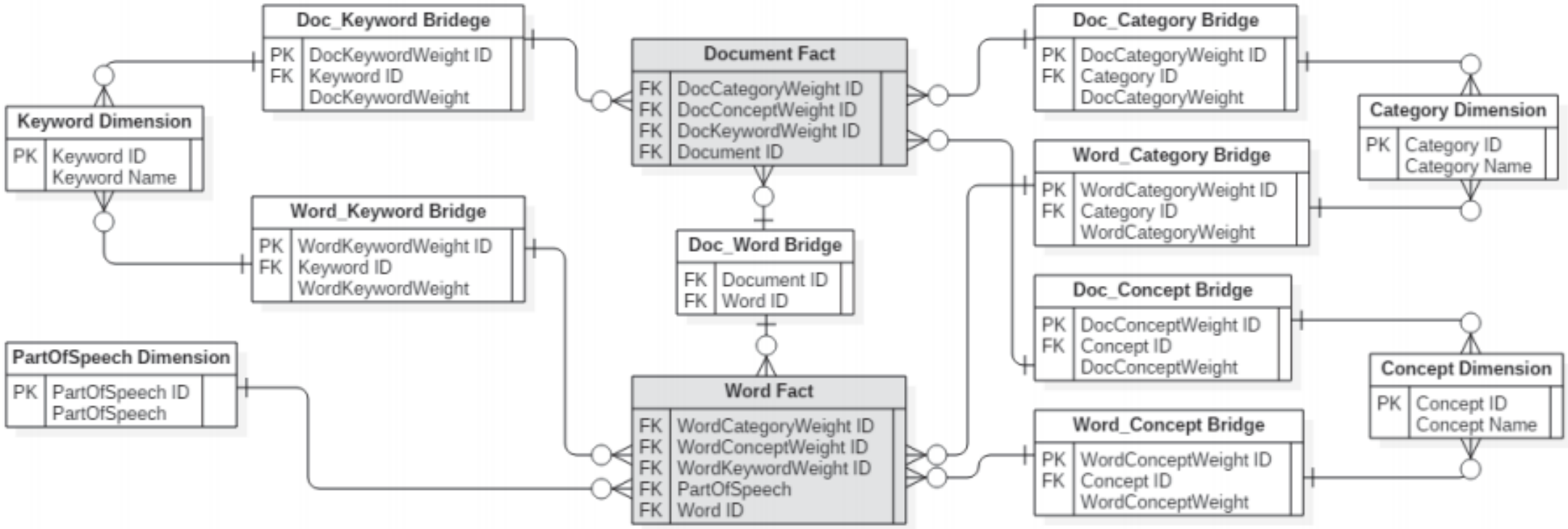


Fig. 12. An example star schema for journal paper warehousing.

Grain: document or term ?

Text Warehouse 예




Doc_Keyword Bridege
PK DocKeywordWeight ID
FK Keyword ID
DocKeywordWeight
Keyword Dimension
PK Keyword ID
Keyword Name
Word_Keyword Bridge
PK WordKeywordWeight ID
FK Keyword ID
WordKeywordWeight
PartOfSpeech Dimension
PK PartOfSpeech ID
PartOfSpeech
Document Fact
FK DocCategoryWeight ID
FK DocConceptWeight ID
FK DocKeywordWeight ID
FK Document ID
Doc_Word Bridge
FK Document ID
FK Word ID
Word Fact
FK WordCategoryWeight ID
FK WordConceptWeight ID
FK WordKeywordWeight ID
FK PartOfSpeech
FK Word ID
Doc_Category Bridge
PK DocCategoryWeight ID
FK Category ID
DocCategoryWeight
Category Dimension
PK Category ID
Category Name
Word_Category Bridge
PK WordCategoryWeight ID
FK Category ID
WordCategoryWeight
Doc_Concept Bridge
PK DocConceptWeight ID
FK Concept ID
DocConceptWeight
Concept Dimension
PK Concept ID
Concept Name
Word_Concept Bridge
PK WordConceptWeight ID
FK Concept ID
WordConceptWeight

Context definition in Text Warehousing

- Document context:
  - contextual factors of the analyzed documents
  - topics, domains, metadata such as title, author, date, and etc
- User context:
  - user profile which is based on his query log containing the history of queries
  - user preferences for the analysis dimensions.

Issues

- Schema design
  - Grain
  - Textual dimensions
  - Semantic dimensions
  - Bridge tables
- Text understanding
  - Concept extraction for documents and terms
  - Concept hierarchy
  - Weighting for grains

ConteXtual Text cube model: CXT-Cube

- Document representation
  - Term vector
  - d = <w_{t1}, w_{t2}, w_{t3}, … , w_{t1}>
- Measure definition
  - Represents each document d by several vectors of weighted concepts

$$M = < \overrightarrow{d_{Dim_1}}, \overrightarrow{d_{Dim_2}}, \ldots, \overrightarrow{d_{Dim_*}} >$$

$$\overrightarrow{d_{Dim_r}} = < w_{c_1}, w_{c_2}, \ldots, w_{c_n} >$$

Text Measures

- Measure definition

$$\overrightarrow{d_{Dim_r}} =< w_{c_1}, w_{c_2}, \ldots, w_{c_n} >$$

  - the vector of weighted concepts of a document *d* in a vector space dimension specific to Dim_r and w_{ci} is the weight assigned to the concept c_i.

- Example

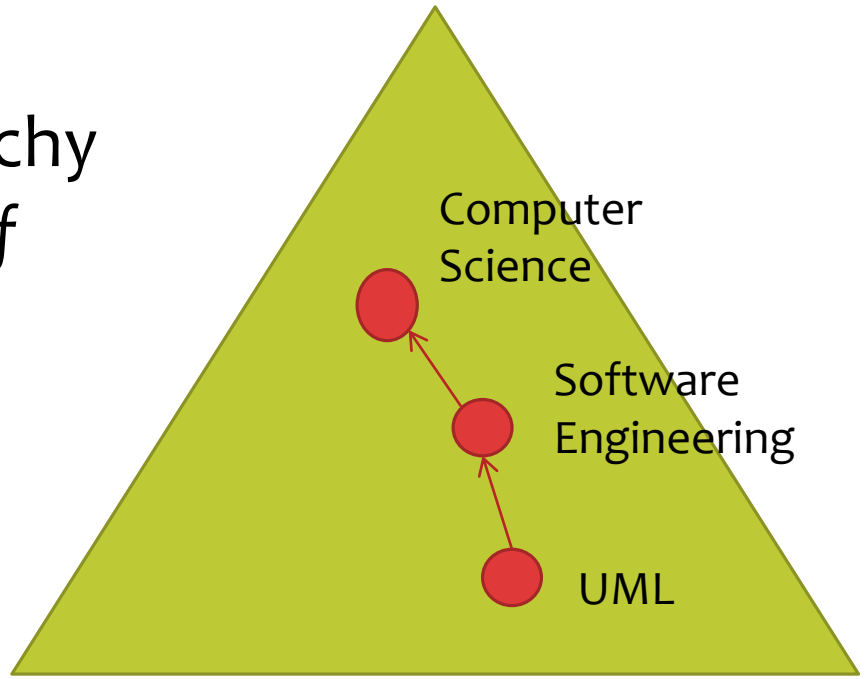
$$M =< \overrightarrow{d_{dim_z}}, \overrightarrow{d_{Dim_l}}, \overrightarrow{d_{Dim_t}} >$$

$\overrightarrow{d_{dim_z}}, \overrightarrow{d_{Dim_l}}, \overrightarrow{d_{Dim_t}}$: vectors of dimensions: TOPIC (Dim_z), LOCATION (Dim_l), and TIME (Dim_t)

Relevance Propagation

- Step 1
  - compute the weights of the document terms which exist in the concept hierarchy by using the method Term Frequency *Tf* computed by the following formula:

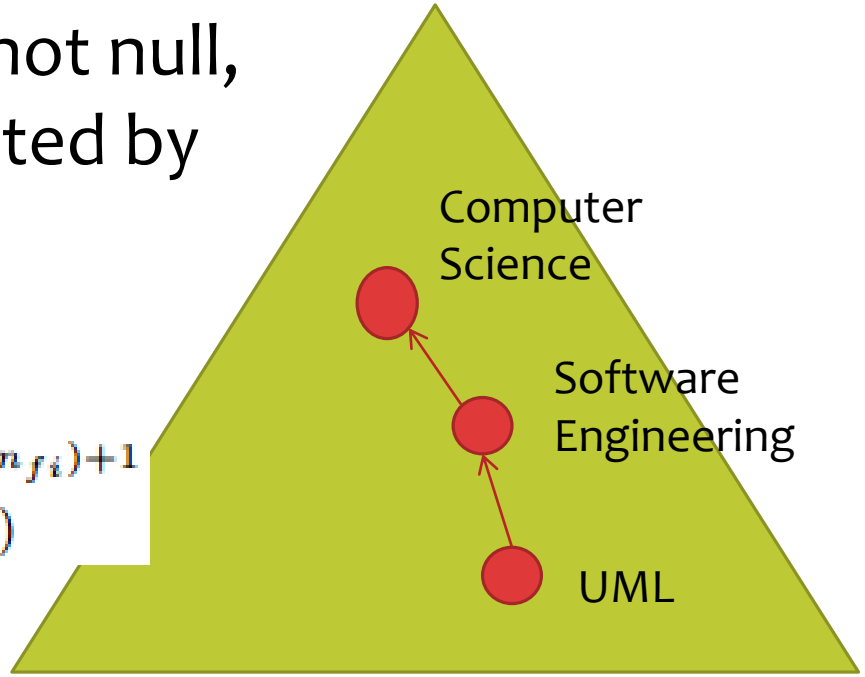
$$Tf_{t,d} = \frac{n_{t,d}}{N_d}$$



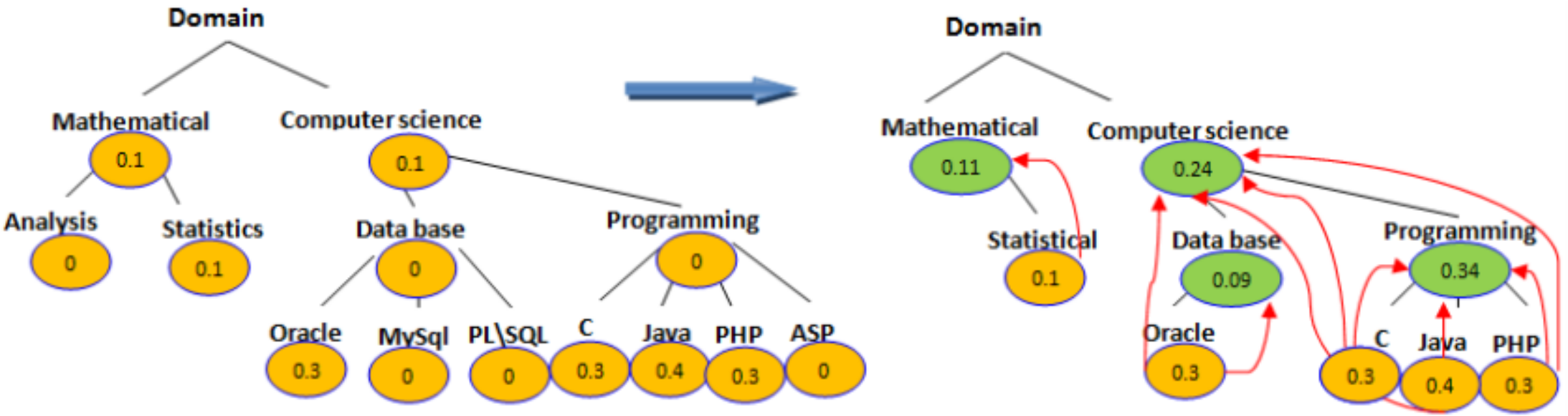
Relevance Propagation

- Step 2
  - For each leaf node which has a weight not null, the weights of his ancestors are computed by the formula below:

$$Weight(n_k, n_{fi}) = Weight(n_k) + Weight(n_{fi})^{distance(n_k, n_{fi})+1} \quad (2)$$



Relevance Propagation



more important

Result: $\overrightarrow{d_{Dim_z}} = computer{-}science(0.24), programming(0.34), Java(0.4), php(0.3), c(0.3), database(0.09), oracle(0.3), math(0.11), statistics(0.1).$

Text-OLAP based IR

$$Sim(d,q) = \cos a = \frac{\sum_i w_{ti} * wq_{ti}}{\sqrt{\sum_i w_{ti}^2 * \sum_i wq_{ti}^2}}$$

$$Sim(d,q) = \sum_{i=1}^{n} (\alpha_i \times Sim(d_{Dim_i}, q_i))$$

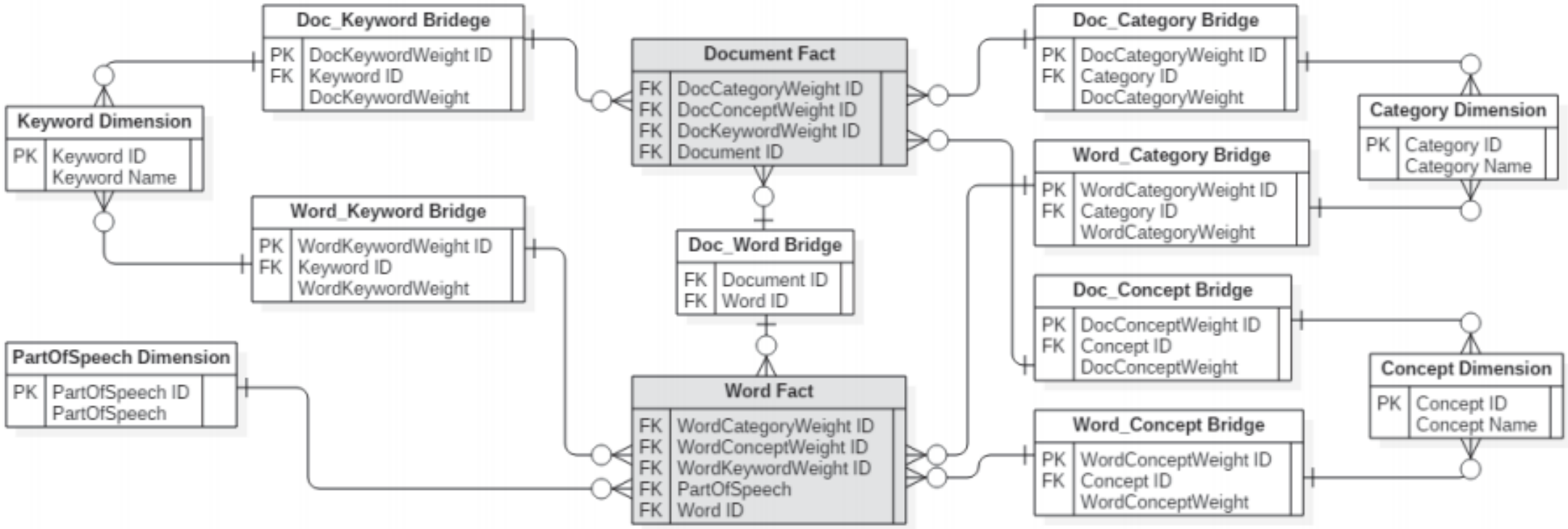
$$Sim(d,p) = \sum_{i=1}^{n} (\alpha_i \times Sim(d_{Dim_i}, p_i))$$

$$ORank(d) = \beta \times Sim(d,q) + \gamma \times Sim(d,p)$$

CXT-Cube

- For CV data
  - Fact table: CV documents
  - Dimension tables
    - Semantic dimensions: TOPIC, LOCATION
      - TOPIC : hierarchy of domain skills, ex) data science – computer science - science
      - LOCATION : ex) city-region-country

A Star Schema for Text Warehouse



DW-based Machine Learning