VIDYALANKAR INSTITUTE OF TECHNOLOGY

**Identifying Relationships for Entity Relationship Query**

SUBMITTED BY

**NEHA TEMBE – 12106B0026  
AHMED SABEEH – 12106B0035  
DHRUV KHARWAR – 12106B0068**

UNDER THE GUIDANCE OF

**PROF. PANKAJ VANWARI  
 DEPARTMENT OF COMPUTER ENGINEERING**

ABSTRACT

The information available on the web, with regard to certain entities is in latent, unstructured form. Hence, the main aim of our system, using information extraction (IE), is to understand the semantic equivalence of phrases and map them to a canonical form. The patterns will be organized into synonyms and subsumptions.

**Example :**  
  
The fact that Natalie Portman has won the Oscar should be recognized as an instance of the hasWonPrize relation (with type signature person \_ award), but it occurs in very diverse forms: “Portman was honoured with an Oscar”, “. . . received. . . ”, “. . . thanked the academy for . . .”, “Awards received: Academy Award 2011” (in a Web table or list), “Prizes: Oscar for Best Actress”, etc. In addition, a paraphrase like “received” could also denote different relations such as almaMater (“received her Ph.D. degree from”), critizedBy (“not well received by”), or meets (“received by the queen”).

Our system will deal with the issue of comprehensive gathering and systematically organizing patterns for an open set of relations. The system can then be queried to show a relationship between entities. The system will be able to to detect and disambiguate named entities in text and extract binary relations between entities based on patterns in textual or semi structured contents.

This system can be used to organize a large number of relational patterns into sets of synonymous patterns and finally into a hierarchy of entities. It can boost information extraction (IE) and knowledge based population tasked by means of its repository of paraphrases for relations. It also enables advanced search over “subject-predicate-object” data.

PROBLEM DEFINITION

* Our main aim is to detect and disambiguate named entities in text and extract binary relations between entities based on patterns in textual or semi structured contents.
* Collection of semantically-typed relational patterns.
* The patterns are organized into synonyms and subsumptions.
* Joins entities which are related to each other by mapping them.

eg. Sources may use the verbal phrases “received” or “was honored with” to say that person won an award.

* Also , we make an effort to systematically harvest textual patterns from text corpora, to group similar patterns into sets, resulting into a subsumption hierarchy.
* In order to acquire the goals mentioned above, we make use of dictionary of entity-class pairs   
  eg. Provided by knowledge bases like Freebase, or Dbpedia.
* For extracting relations, we refer wikipedia data.

This system can be used to organize a large number of relational patterns into sets of synonymous patterns and finally into a hierarchy of entities. It can boost information extraction (IE) and knowledge based population tasked by means of its repository of paraphrases for relations. It also enables advanced search over “subject-predicate-object” data.

LIST OF REFERENCES

* <https://people.mpi-inf.mpg.de/~weikum/index.html#publications>
* Max-Planck-Institut für Informatik: YAGO-NAGA <http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/>
* Max-Planck-Institut für Informatik: Software <http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/software/>
* <http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/patty/>
* <https://d5gate.ag5.mpi-sb.mpg.de/pattyweb/>
* DBpedia
* Freebase
* Black book of the project based on past work

BRIEF REPORT ON LITERATURE SURVEY

We are going to refer to last year’s BE project of Miss. Bhakti Hinduja and group. They have used the concept of Sense-making which focuses on making sense of ambiguous contexts and continuously making the found knowledge more precise based on disambiguating the context. The used effective analysis tools to find the key entities and their relations in the sense making task. Their most relation identification work focused on the relations like ‘’is-a’’ or ‘’part-of’’ , which expresses the connections between entities in a hierarchical structure. However relations in all environments need not be limited to hierarchical relations; this is what they stated. For example, in the field of customer relationship management, it is important to capture the relations between a company and its product, and a product and customer reviews. Then in marketing and business intelligence area, it is important to identify ‘important’ relations based on extracted entities. To help with these goals, they proposed a system to understand the phrase relation-pattern categorizing them in various category definitions.

To implement this, they used machine learning as the main area of interest. Also they did a brief research in “Knowledge Discovery in Text Mining Technique Using Association’’.

BRIEF REPORT ON LITERATURE SURVEY

* Using the previous year’s project information, we aim to detect and disambiguate named entities in text and extract binary relations between entities based on patterns in textual or semi structured contents. Also we make an effort to systematically harvest textual patterns from text corpora, to group similar patterns into sets, resulting into a subsumption hierarchy.

The applications of the project are as follows-

* To organize a huge number of relational patterns into sets of synonymous patterns and finally into a hierarchy.
* It can boost Information Extraction and knowledge based population tasks by its repository of paraphrases for the relations.
* It can improve open Information Extraction by associating type signatures with patterns.
* Help discovering “Web Witnesses” when accessing the truthfulness of search results or statements in social media.
* It enables advanced search over subject -predicate -object data.
* It will be of interest to Database community.

PROJECT ANALYSIS

**Hardware requirements:**

* 64-bit processor
* 8 GB RAM
* 80 GB hard-disk space (minimum, preferably SSD)

**Software requirements:**

* Linux 64-bit operating system
* Hadoop
* PostgreSQL
* TagMe
* OpenNLP

**Feasibility**

The assessment is based on an outline design of system requirements, to determine whether there is technical expertise to handle completion of the project. The concern is whether the proposal is technically feasible.

**Operational feasibility**- Operational feasibility is a measure of how well a proposed system solves the problems and how it satisfies the requirements analysis phase of system development. Using our approach and concept, there is no loss of information as also, it can improve open Information Extraction by associating type signatures with patterns.

**Economic feasibility-** This involves questions such as how much time is available to build the new system and amount of resources and tools required, as mentioned in the hardware and software requirements. All the tools are available and hence our project is feasible in terms of the resources required.

**Cost Estimation:**

* As per the hardware requirements, we need 8GB RAM. Normal availability of RAM is 6GB, so we need another 2 GB RAM which will cost around Rs. 1000 to Rs. 1500.
* Also, we need sdd and hdd which will cost around Rs. 8000.

**Timeline:** (Make gantt chart for the entire project development life)

PROJECT DESIGN

**DATA FLOW DIAGRAM LEVEL 0**

**C:\Users\neha tembe\Downloads\DFD - Level 0.png**

**DATA FLOW DIAGRAM LEVEL 1**

C:\Users\neha tembe\Downloads\Level-1.png