**Identifying Relationships for Entity Relationship Query**

***Synopsis Report submitted in partial fulfillment***

***of the requirement for the degree of***

**B. E.(Computer Engineering)**

Submitted By

NEHA TEMBE  
AHMED SABEEH  
DHRUV KHARWAR

Under the Guidance of

Prof. PANKAJ VANWARI

Department of Computer Engineering



Vidyalankar Institute of Technology

Wadala(E), Mumbai 400 037

University of Mumbai

2014-15

### 

### CERTIFICATE OF APPROVAL

**For**

**Project Synopsis**

This is to certify that

**NEHA TEMBE  
AHMED SABEEH  
DHRUV KHARWAR**

Have successfully carried out Project work entitled

**Identifying Relationships for Entity Relationship Query**

in partial fulfillment of degree course in

Computer Engineering

As laid down by University of Mumbai during the academic year

2014-15

Under the Guidance of

Prof. Pankaj Vanwari

Signature of Guide Head of Department

Examiner 1 Examiner 2 Principal

### DECLARATION

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Name of student Roll No. Signature

1. NEHA TEMBE 12106B0026

2. AHMED SABEEH 12106B0035

3. DHRUV KHARWAR 12106B0068

Date: 26/10/2015

### ACKNOWLEDGEMENT

On the very outset of this project synopsis, we would like to extend our sincere and heartfelt obligation towards all the people who have helped the three of us in this endeavour. Without their active guidance, cooperation and encouragement, we would not have made headway in the project. We are ineffably indebted to Prof. Pankaj Vanwari for conscientious guidance through the insights, inputs and the encouragement to accomplish this project.

I extend my gratitude towards Vidyalankar Institute of Technology for giving me this opportunity.

Thanking You,  
   
Neha Tembe

Ahmed Sabeeh

Dhruv Kharwar

Table of Contents

[CERTIFICATE OF APPROVAL 2](#_Toc433551844)

[DECLARATION 3](#_Toc433551845)

[ACKNOWLEDGEMENT 4](#_Toc433551846)

[ABSTRACT 6](#_Toc433551847)

[AIMS AND OBJECTIVES 7](#_Toc433551848)

[LITERATURE SURVEY 8](#_Toc433551849)

[PROBLEM STATEMENT 9](#_Toc433551850)

[SCOPE OF THE PROJECT 10](#_Toc433551851)

[PROPOSED SYSTEM 13](#_Toc433551852)

[METHODOLOGY 14](#_Toc433551853)

[ANALYSIS 16](#_Toc433551854)

[USE CASE DIAGRAM 17](#_Toc433551855)

[REFERENCES 18](#_Toc433551856)

[PROJECT DESIGN 19](#_Toc433551857)

[HARDWARE AND SOFTWARE REQUIREMENTS 20](#_Toc433551858)

## 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 :**  
  
If we enter “Dhoni plays for a team’’ , ‘’Dhoni is a member of a team’’ or ‘’Dhoni play team’’ , then in all these cases, our system will analyse the terms ‘’plays for’’ ‘’is a member’’ and ‘’play’’ and understand that they come under the same category and mean the same, and eventually, it will display the same result for all the three statements, where “is a member” falls under a broader category and might possibly give more results.

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.

## AIMS AND OBJECTIVES

* Our main aim is to map predicates to phrases in knowledge base. It identifies relationships for entity relationship query.
* 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.

## LITERATURE SURVEY

* We are going to refer [1]. 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.
* Using [1] 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.

## PROBLEM STATEMENT

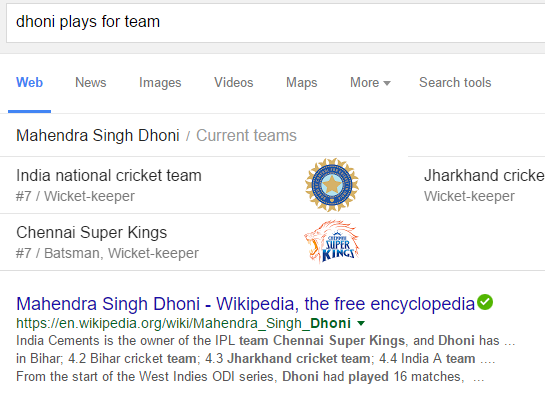
* Our main aim is to detect and disambiguate named entities in text and extract binary relations between entities based on patterns in textual or semi structured contents.
* Collection of semantically-typed relational patterns.
* The patterns are organized into synonyms and subsumptions.
* Joins entities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

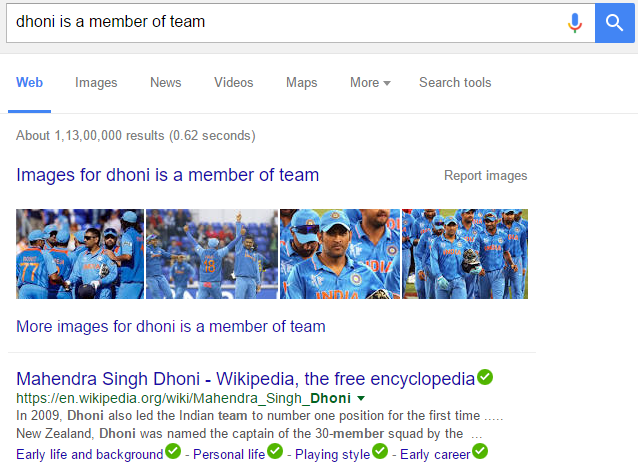
* Also , we make an effort to systematically harvest textual patterns from text corpora, to group similar patterns into sets, resulting into a subsumption hierarchy.
* In order to acquire the goals mentioned above, we make use of dictionary of entity-class pairs   
  eg. Provided by knowledge bases like Freebase [2], or Dbpedia [3]
* For extracting relations, we refer wikipedia data.

## SCOPE OF THE PROJECT

* The main aim of our project is to map the predicates with phrases in knowledge base. This will in turn, help us to ameliorate the searching process.
* In the existing system, if we enter ‘’Dhoni plays for a team’’, it will give the following result.



* Now, if we enter ‘’Dhoni is a member of a team’’ , it will display:



Considering the above example, our system will give the same result for both the statements. That is, even if we enter ‘’Dhoni is a member of a team’’, then it will understand the relation and give the same result as that of the previously entered statement.

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 Information Extraction by associating type signatures with patterns.
* It enables advanced search over subject -predicate -object data.
* It will be of interest to Database community.

## PROPOSED SYSTEM

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

Take input from the user and give search result.

The system will take the user query or sentence and will generate the search result by processing the parts of the query, identifying the relation and mapping it to the correct categories as intended in the query or as processed by the system.

**SUBJECT** 🡪 Dhoni plays for a team 🡨 **OBJECT**

**RELATION**

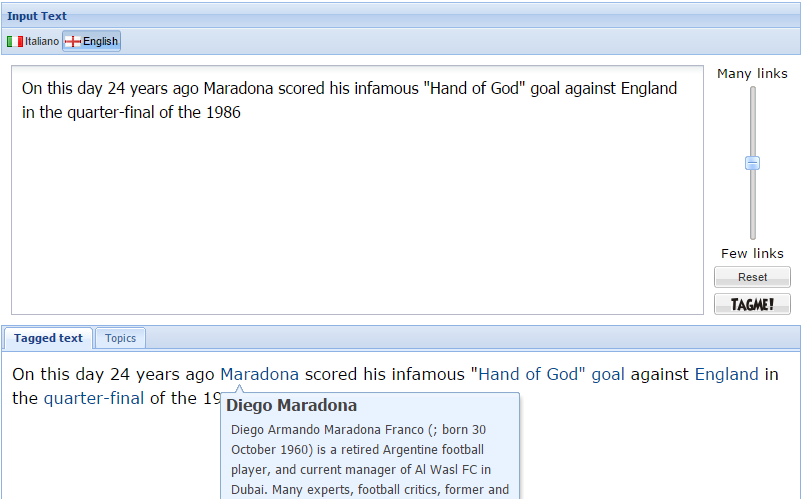
We have referred to one of the projects of The Max Plank Institute, that is, PATTY: A large resources of relational patterns. It is a collection of semantically-typed relational patterns mined from large corpora. The patterns are organised into synonyms and categories.

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.

## METHODOLOGY

* To start with , we are going to use the following tool:
* TagMe tool

TagMe: is a powerful tool that is able to identify on-the-fly meaningful short-phrases (called "spots") in an unstructured text and link them to a pertinent Wikipedia page in a fast and effective way. This annotation process has implications which go far beyond the enrichment of the text with explanatory links because it concerns with the contextualization and, in some way, the understanding of the text.



With reference to Dhoni example,



* The next step is to categorize the entities. For this purpose, we generate the entity index and the category index.
* We map the relations by processing the Knowledge base and the data obtained from Wikipedia.
* These two that is entity index, category index and the relation phrase is then processed by system to obtain the final result.

## ANALYSIS

**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 Information Extraction. The current system is planned by taking into consideration the requirements needed to implement the system so that it can be successfully completed in the given amount of time.

**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 a SSD which will cost around Rs. 8000.

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

## USE CASE DIAGRAM

**C:\Users\neha tembe\Downloads\UseCase.png**

# **REFERENCES**

|  |  |
| --- | --- |
| [1] | B. Hinduja, Y. Kumavat and M. Sambe, Extraction of knowledge from text, Mumbai, 2014. |
| [2] | “Freebase,” [Online]. Available: http://www.freebase.com/. |
| [3] | “DBpedia,” [Online]. Available: http://wiki.dbpedia.org/. |
| [4] | N. Nakashole, G. Weikum and . F. Suchanek, “PATTY: A Large Resource of Relational Patterns,” [Online]. Available: http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/patty/. |
| [5] | Max-Planck-Institut für Informatik “YAGO-NAGA,” [Online]. Available: http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/. |

* <https://people.mpi-inf.mpg.de/~weikum/index.html#publications>
* Max-Planck-Institut für Informatik: Software <http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/software/>
* <https://d5gate.ag5.mpi-sb.mpg.de/pattyweb/>

## 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**

## HARDWARE AND SOFTWARE REQUIREMENTS

**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(Explain)
* OpenNLP

**IMPLEMENTATION PLAN FOR NEXT YEAR**

|  |  |  |  |
| --- | --- | --- | --- |
|  | DURATION | START | FINISH |
| 1.Designing the system | 40 days | 11/20/2015 | 12/11/2015 |
| 2. Coding | 30 days | 1/1/2016 | 2/1/2016 |
| -Creating prototypes |  |  |  |
| -Basic Implementation |  |  |  |
| -Further development |  |  |  |
| 3.Testing | 10 days | 2/3/2016 | 2/13/2016 |
| -Unit testing |  |  |  |
| -Integration testing |  |  |  |
| -System testing |  |  |  |
| -Acceptance testing |  |  |  |
| 4.Maintenance | 20 days | 2/14/2016 | 3/04/2015 |
| -Bug removal |  |  |  |