A SEMANTIC RETRIEVAL SYSTEM FOR EXTRACTING RELATIONSHIPS FROM BIOLOGICAL CORPUS

Hassan Mahmoud, Saif Salah Kareem and Tarek El-Shishtawy

Department of Information Systems, Faculty of Computers and Informatics,
Benha University, Egypt.

ABSTRACT

The World Wide Web holds a large size of different information. Sometimes while searching the World Wide Web, users always do not gain the type of information they expect. In the subject of information extraction, extracting semantic relationships between terms from documents become a challenge. This paper proposes a system helps in retrieving documents based on the query expansion and tackles the extracting of semantic relationships from biological documents. This system retrieved documents that are relevant to the input terms then it extracts the existence of a relationship. In this system, we use Boolean model and the pattern recognition which helps in determining the relevant documents and determining the place of the relationship in the biological document. The system constructs a term-relation table that accelerates the relation extracting part. The proposed method offers another usage of the system so the researchers can use it to figure out the relationship between two biological terms through the available information in the biological documents. Also for the retrieved documents, the system measures the percentage of the precision and recall.

KEYWORDS

Inverted list, information retrieval, Gene Ontology, Information extraction, Relationship extraction, Pattern recognition.

1. INTRODUCTION

Nowadays there is a tremendous growth of biological experimental data and textual information. So accessing biological data become a challenge. Now in biomedical digital libraries, the need of information retrieval system become a must. Using of information retrieval (IR) helps in retrieving documents based on users needs. Ontology always defined as a regular conceptualization of a specific domain to be understandable by humans and machine-readable.

Gene Ontology is ontology of the various biological ontologies and at the same time is most valuable bio-ontology. Gene Ontology will be used in constructing the proposed system presented in this paper. Information extraction is used to extract useful information in a structured form from not structured documents or semi-structured documents. Besides the endless increase of information, the importance of biological extraction methods is also increased. Knowing the relations between biological words by extracting it from a document helps the users to know the relative that group these words.

Relationship Extraction works with the challenge of getting associations exists between terms within a phrase of text. Common methods for relationship extraction are always apply co-occurrence based [1], rule-based [2], and kernel-based [3] techniques. In the medical field,

DOI:10.5121/ijcsit.2018.10104

relationship extraction is used to define the relationships between proteins [2, 4, 5]. Pattern recognition belongs the subject of machine learning and it focuses on regularizing data and recognizing patterns [6].

The pattern recognition implemented systems are trained from labeled data and this is called supervised learning, but on the other side in unsupervised learning, algorithms used to observe unknown patterns when no labeled data are available [6]. Pattern recognition always categorized based on the type of learning procedure that is used to construct outputs. Supervised learning works by training the system by a set of pre-defined words that have been properly grouped and assigned to the correct and accurate output. Then a learning method constructs a template that tries to gather 2 objectives have a conflict [7].

The system proposed in this research infers semantically related words uses Gene Ontology. It takes the t words as an input then it retrieves documents which is relevant and contain the words entered in the input and its parents or synonym. It also constructs a term - relation table that accelerate the relation extracting part. The presented system offers another usage of the system so the researchers can use it to figure out the relationship between two biological terms through the available information in the biological documents. It also measures the percentage of the recall results and precision results of the resulted outputs. Unsupervised learning inherent new patterns from the data to be used in determining the correct output for new data instances [7].

The research is written on this way: Samples of the other researchers work that belongs to the subject of the paper is written in section number 2. In section number 3, the materials and methods used in the presented system are described and explained. In section number 4, experiments of work are illustrated. Discussions and results are mentioned in section number 5, then future work and conclusion are drowned in section number 6.

2. RELATED WORK

Previous works were studied for information retrieval systems subject and extracting semantic relationships.

Dr AK Sharma., [8] showed that the system can retrieve relevant document belong to a certain context with ignoring others contexts. The context-based index used to retrieves information based on context, not keywords. This helps in developing the inputs quality. The idea of the paper is very useful since it improves the retrieving of information. It allows the user to have only the documents that contain his search information in a certain context. Form the other side adding the additional term to the search information may not be useful in certain domain since I know that this information does not suffer from the polysemy problem.

Meng et al., [9] present a system that can filter a huge amount of irrelevant documents. The authors mentioned that the user searches a set of words by exact keywords matching after that a group of documents is retrieved for this search query. Terms are weighted based on Gene Ontology and this benefits in clustering the document based on considering the hierarchy of biological words. The authors showed that the users' feedback is a very small number of representative documents. Also, authors proposed a prototype for biomedical literature search system based on this iterative search paradigm. They concluded that collecting the user feedback to get the relevant document must be retrieved in not a good approach since the user should review and rate each retrieved document manually.

Silvestri et al., [10] proposed an algorithm that used a k-means-like algorithm for clustering. They use it to compute reordering effectively with time complexities and linear space. Their results

conducted with an improvement up to 23% roughly in the rate of compression and this is achieved using a real test collection, resulted in. In this paper the author focused on the way to save the inverted list and resulted from an improved rate of compression achieved, the k-mean clustering has a problem that it is computationally difficult (NP-hard).

Huang et al., [11] showed that their system is useful for extract knowledge from a large amount of MEDLINE abstracts. The is very useful since it merges many different ontologies to develop the structure and organization of concept. The system applies a two-level pattern learning algorithm to construct patterns which regulated in a hierarchy. Also, the system innovates a weighted matching model to balance the coverage and accuracy of the system.

Mangla et al., [12] developed an index based retrieval system. They proposed comparing and multiplication method that helps in giving the relevant result. Authors used stemming lexical form words. The mentioned that the length of the index table is varied for different documents based on the threshold value. They reduce the space taken by the system by removing the stop words and they mentioned that it is not helpful. The major objective of presented work is to generate the contextual model for retrieving relevant documents.

Nebhi et al., [13] developed an efficient Ontology-based system for Twitter Information Extraction. Authors provided an integrated cleared method its core is the syntax similarity and popularity score. The results show that the system performance significantly better-using process for disambiguation. To evaluate the system the authors used the processing resources on the evaluation corpora of 116 short messages from BBC News, The Times Twitter accounts, and New York Times. They annotated these documents manually with the use of the DBpedia ontology. They use the person, location, and organization named entity categories. They improved the extraction efficiency of the system and the traditional F-measure become 86% and augmented F-Measure become 90%.

Buscaldi et al., [14] presented an open-source free Lucene(YaSemIR) based system for semantic information retrieval. The system merges ontologies with OWL format and maps them to the body of terms connected with every ontology to semantically index a text collection. The authors annotate the concepts in documents and the ontology to collect the meta information to expand them. YaSemIR can work with different ontologies, on different kinds of documents.

Mihalcea et al., [15] developed a rule-based method for the semantic disambiguation and recognition in tweets with the entities that named. The authors developed the system depending on understanding the meaning of the word. The system identified the critical named entities like organization, person and links the information by general ontology, namely WordNet. They regarded the WS model to the long-term target of Semantic Web creation. They proposed "Semantics Word Model", which is more simple than RDF or XML, but it has a great advantage of being available with other resources and already exists methods.

3. MATERIALS AND METHEDS

The proposed system developed to retrieve biological documents and extract biological relationships from biological documents. The aim of the system is to semantically retrieves relevant documents based on the input terms and construct term-relation table that contains two biological terms and the relation between them. Also, the aim of the system is to classify the retrieved documents into several classes.

3.1. Research objectives

1. Retrieving information efficiently, in a high quality, and efficiently while ensuring the retrieval of all files containing the words, the words synonyms and words that are collected by father relationship.

2. Ordering and separating the files and dividing them into groups. The files are arranged according to their importance to the user. The order of files as follows :

   - files containing the searched words
   - words containing the searched words and synonyms
   - words containing the searched words and words collected by the relationship father).

3. Collect all terms information and save it in the ontology database.

4. Point #3 has various benefits such as :

   - Easy to store information and easy retrieval
   - Make synonyms retrieval easier
   - Easily retrieve the relationship between terms
   - Improve the retrieving process quality because of using ontology.

5. Retrieve and restore relations that exists between words.

6. Inferring new relationships.

7. Improves the time taken in files retrieving process.

3.2. Steps used to implement the system

Figure number 1 shows the steps used to develop the proposed system.

As algorithm in figure 1, ontology was build based on the gene ontology file and saved in the database called "ontology". Inverted list and position list from the standard Boolean model were built based on the documents exist in the corpus. Ranking and classification part were implemented. The rank value of the document is calculated based on the occurrence of the words entered in user's query in this document.The system classify the documents to three classes, they are:

- **Class 1**: contains all document that contain the same phrase of the searched words.
- **Class 2**: contains all document that contain the synonyms of the searched words.
- **Class 3**: contains all documents that contain the childs (from database) of the searched words.

The classification part allow the user to have all the information needed about the searched words so he will get all the documents that contain the same phrase of the searched words, the synonyms of the searched words and the childs of the searched words. As shown in figure 1, the query expansion part was implemented based on 3 levels:

- **Exact phrase of words**: get all document that contain the same phrase of the searched words

- **synonyms of words**: get all document that contain the synonyms of the searched words
- **is_a relation**: get all documents that contain the Childs (from database) of the searched words.

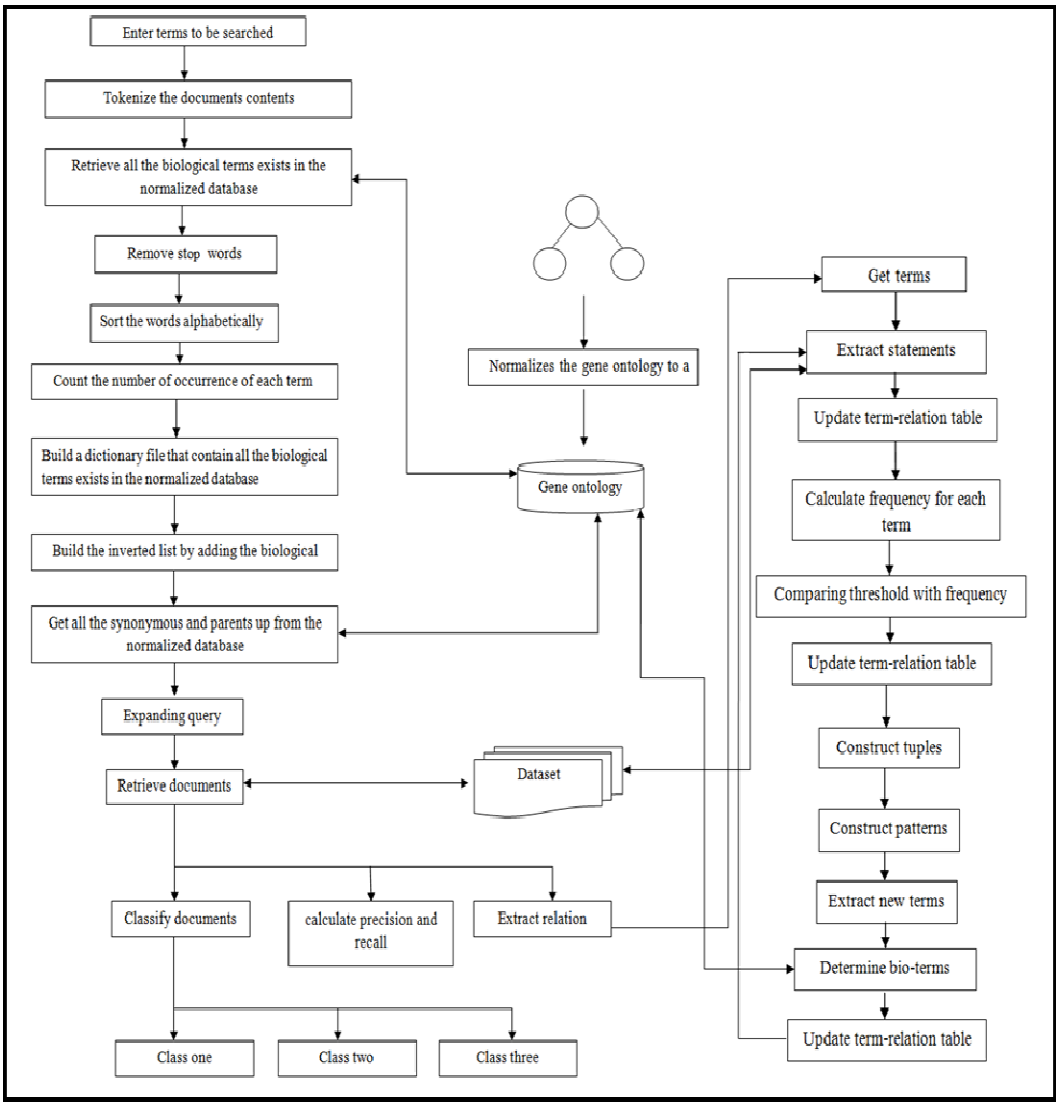

Figure 1. Implemented system steps

3.3. Term – relation table

As shown in Figure 1, to build this table we made the following steps:

3.3.1- Initiate term – relation table: We begin with the two terms CEBPG and DNA

T1	T2	Relation - Document
CEBPG	DNA	?

Results of this step:

The following statements are samples of the retrieved statements (terms are underlined and the relation has a bold effect):

Of appropriate key is the absence of correspondence of the spot genes with CEBPG, which **binds** to the equivalent realization position as CEBPG that heritage notice section within each of the antioxidant or DNA patch genes

CEBPG is a trimmed CEBPG transcription agent [19] and maintains the orders required for DNA binding and heterodimer structure, but requires the orders needed for transactivation [20]

Conclusion We decide that in non-Before Christ individuals, CEBPG **regulates** transcription of key antioxidant or DNA repair genes in NBEC and that in smokers who Produce CEBPG **regulation** is sub-optimal for enough amount of antioxidant or/and DNA genesrepair to effect raised risk

3.3.2- Building patterns

This step contains certain number of instructions used to construct patterns. Patterns will help in extracting existing terms from the documents

3.3.2.1- Calculate frequency

To determine which word between the two biological terms is a biological relation we must count the occurrence of all words exist between the two terms. After counting the occurrence we determine a threshold so any word had an occurrence value greater than the threshold value can be categorized as a relation.

3.3.2.2- Determining threshold

The threshold value was determined from calculating the average of terms frequencies resulted from 100 running experiments. The resulted threshold value was "20". Form the pervious results in step 1 (Initiate term – relation table), the occurrence of the two relations "Bind" and "Regulate" Exceeded the value of the threshold. So "Bind" and "Regulate" have been considered as biological relations between the two terms CEBPG and DNA.

Now update the table:

T1	T2	Relation - Document
CEBPG	DNA	([Bind, Regulate], 2)

3.4- Patterns construction

After finding the relation between "CEBPG" and "DNA", we begin to build a tuples from the retrieved statements. The tuple consist of [prefix, term 1, relation, term2, and suffix]. Both prefix and suffix have a length of 7 characters. Prefix is the seven characters before term1 in the retrieved statement and suffix is the seven characters after term 2 in the retrieved statement. From the previous result in step 1 we can end up with the following tuples.

Tuples:

['site as', CEBPG, binds , DNA, ' repair']

['iduals,', CEBPG, regulate , DNA, 'repair']

[' lop BC ', CEBPG, regulate , DNA, ' repair']

After that we fetch about the similarities between tuples to construct patterns. From the previous tuples we consider that tuple 2 and 3 are similar because they have the same relation and suffix so we can end with the following patterns.

Pattern:

< (.+?), (.+?), binds , (.+?), (.+?) >

< (.+?), (.+?), regulates, (.+?),(.+?)>

The constructed patterns can be used to extract another biological terms that have relations "Bind" and "Regulate" between each other.

Results of this step: After applying these patterns, the following statements are samples of the retrieved statements:

Sam 68 RNA Binding Protein Guards Maice from Age-Related Bone hurt Abstract The source substrate related in mitosis of 68 kDa (Sam68) is a KH-type RNA binding protein that has been presented to **regulate** various looks of RNA metabolism; nevertheless, its physiologic role has outlived obscure.

Of these, 2,118 eQTLs were placed within 20 Mb (roughly 10 cM) of the similar gene, acceptable describing **eQTLs** regulated by **cis-acting** variation within the gene itself **ESG1** was first recognized as a reproduction Ph34 that was down-**regulated** by retinoic acid in embryonic carcinoma cells [13]

We conclude :non-Before Christ individuals, CEBPG**regulates** transcription of key antioxidant or DNA repair genes in NBEC and that in smokers thatproduce BC, CEBPG regulation is sub-optimal for a enough number of antioxidant and/or DNA repair genes to produce raised danger
Of special note is the lack of correlation of the destination genes with CEBPG, which **binds** to the same recognition site as CEBPG that shares recognition site within each of the antioxidant or DNA repair genes

Apply step number 2 and 3 again. Now update term – relation table:

T1	T2	Relation - Document
CEBPG	DNA	([Bind, Regulate], 2)
RNA	RNA metabolism	([Regulate], 4)
eQTLs	cis-acting	([Regulate], 6)
ESG1	retinoic acid	([Regulate], 8)
CEBPG	antioxidant	([Bind, Regulate], 2)

These steps will be repeated until fining the patterns come over again. Now the system has a term- relation table so the relation between words can be extracted easily from it.

4. MATERIALS AND METHEDS

We have concluded 2 experiments to evaluate the implemented algorithm. The first experiment compares the precision and recall with work done in [16], while the second experiment compares

the build and extraction with other systems. The following subsection gives the detail of the experiment.

4.1. Dataset Description

Several experiments are executed to ensures the efficiency of the implemented system which retrieves relevant documents and extracts the relationships from the biological documents. The system was tested using 1000 documents contain biological abstracts. The average number of sentences per document is 18 sentences. The average number of words per sentence is 10 words.

4.2. Experiment Procedure

As algorithm in figure 1 to retrieve the documents which contain the terms used on search, our proposed system follow the following steps:

1-Enter the terms you want to search about (terms: biological terms from gene ontology).

2-Get documents form framework and tokenize the contents of the documents and retrieve all the parents and synonymous with the selected search terms from the normalized database.

3-Retrieve every biological keywords exists in the database.

4-Remove stop words from the documents.

5-Sort the words alphabetically.

6-Get the occurrence of alldocuments keywords.

7-Build file with words frequency.

8-Build the inverted list by adding the biological.

9-Expand the search query by fetching the parents of both terms and their synonyms and terms used for the search.

10-Get and retrieve documents that achieve the search query then we generate three classified classed.

11-Calculate precision and recall.

12-Retrieves the relation between terms selected by the user in the query from the ontology.

4.3. Evaluation Procedure

The system measures the recall value and the precision value and displays it.

$$\text{Precision} = \frac{\#\text{ Relevant documents retrieved}}{\#\text{ Retrieved documents}} \quad (1)$$

$$\text{Recall} = \frac{\#\text{ Relevant documents retrieved}}{\#\text{ Relevant documents}} \quad (2)$$

The system calculates the time consumed for the inverted list building and the time of getting back the documents to the user. The system calculates the consumed time by subtracting the system time before and after finishing the task.

5. RESULTS AND DISCUSSION

Because we used the standard Boolean model so there is no chance to retrieves a document that does not contains the terms under search and the percentage of recall and precision equal to 100%.

The performance was improved since that the implemented system retrieves extracted relationship between both words gets in the search query from the biological documents. This helps the researcher to get information about the two entered terms.

According to results, the implemented system retrieves the relevant biological documents in a minimum time. Also, the system extracts all the relationships found between the two terms in the retrieved documents. The relation will be retrieved from term – relation table which minimizes the time of retrieving the relation. The time results table is shown in Table one.

Table One - Time consumed in inverted list building and documents retrieval process.

# documents	Inverted List Building Time in seconds	Document retrieval time in seconds
250	2.65	0.55
500	5.75	0.67
750	10.87	0.78
1000	12.45	0.81
250	2.65	0.55

Shimaa et al. [16] introduced a semi-supervised system to extract binary semantic relations for Arabic text from the Web. The proposed system depended on the pattern-based system. The purpose of the proposed system is to retrieve extracted large list or table from named entities and relations in a specific domain. A small set of a handful of instance relations is required as input from the user. The output is a set of new entities and their relations. The results from four experiments show that precision and recall varies according to relation type.

Table 2 - Comparison between the previous work [16] and our algorithm

Comparison item	Shimaa et al. [16]	Our algorithm
Recall	Max value = 83%	100%
Precision	Max value = 75%	100%
Dependency of results	Depends on entering predefined relations	Depends on threshold value
Inferring new relations	No	Yes
Confidence value	Max value = 0.8	1
Max number of words between inputs	3 word	Any number of words
Type of the text	Structured text	Any type of text

By comparing our system with the system proposed in [16] and as shown in Table 2, our system result on 100% as a percentage of recall and 100% as a percentage of precision and that is

because the system uses the standard boolean model that's retrieves all terms if it exists in the user's query and if it not exist so it will not be retrieved.

Also the system depends on the threshold value that determine which word is the most important and should be retrieved as a relation. On the other side, the system in [16] depends on retrieving the relations based on predefined and pre-entered relations and this can prevent the system from inferring new relation.

In [16] the max value of confidence is 0.8 and this indicates that some uncorrected result can be extracted using the constructed patterns. On the other side our system retrieves data with a confidence value equal to 1 because the system compare the retrieved terms with the biological database to insure that the resulted terms and relations are biological term so all the results are accurate.

6. CONCLUSION

In presented research, we present a technique that successfully retrieves the relevant biological documents and tackles the extracting of semantic relationships from biological documents. The system also constructs a term - relation table that accelerates the relation extracting part. The proposed method offers another usage of the system so the researchers can use it to figure out the relationship between two biological terms through the available information in the biological documents. The measures the values of recall and precision of system documents retrieved.

The method improves semantic performance of the system since it retrieves the relationship between the input words. Term- relation table improves the extracting part since it becomes a reference to all the relationships exist between the biological terms found in the corpus. The proposed method helps the researchers to get more information about the input words by extracting the all relationships between these terms from the biological documents. As a future work, the system can be generalized to e applied in different domains using another datasets and ontologies.

ACKNOWLEDGEMENTS

I need to thank my father and my mother for the efforts they do for supporting me and always being beside me in everything. Also, i need to thank Dr.Tarek El-shishtawy and Dr.Hassan for their efforts with me.

REFERENCES

[1] Ramani AK, Bunescu RC, Mooney RJ, Marcotte EM, Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome, Genome Biol , 6(5), 2001.

[2] Ono T, Hishigaki H, Tanigami A, Takagi T, Automated extraction of information on protein-protein interactions from the biological literature, Bioinformatics , 17(2):155-161, 2001.

[3] Bunescu RC, Mooney RJ, Subsequence Kernels for Relation Extraction, Proceedings of the 19th Conference on Neural Information Processing Systems, 2005.

[4] Blaschke C, Andrade MA, Ouzounis C, Valencia A, Automatic extraction of biological information from scientific text: protein-protein interactions, ProcIntConfIntellSystMolBiol , 60-67, 1999.

[5] Rosario B, Hearst A, Multi-way Relation Classification: Application to Protein-Protein Interaction, Human Language Technology Conference on Empirical Methods in Natural Language Processing, 2005.

[6] Bishop, Christopher M., Pattern Recognition and Machine Learning, Springer. p. vii, 2006.

[7] Jump up, Carvalko, J.R., Preston K, On Determining Optimum Simple Golay Marking Transforms for Binary Image Processing, IEEE Transactions on Computers 21: 1430–33.doi:10.1109/T-C.1972.223519,1972.

[8] Dr AK Sharma. "ontology engine for search." Journal of Computer Applications, (2010).

[9] Hu, Meng, and Jiong Yang. "A System of User-Guided Biological Literature Search Engine." IEEE Data Eng. Bull. 28.4,2005

[10] Silvestri, Fabrizio, RaffaelePerego, and Salvatore Orlando. "Assigning document identifiers to enhance compressibility of web search engines indexes." Proceedings of the 2004 ACM symposium on Applied computing. ACM, 2004.

[11] Huang, M., Zhu, X., Ding, S., Yu, H., & Li, M. (2006, February). ONBRIRES: Ontology-Based Biological Relation Extraction System. In APBC (pp. 327-336), 2006.

[12] Mangla, Neha, and Vinod Jain. "Context based Indexing in Information Retrieval System using BST."

[13] Nebhi, Kamel. "Ontology-based information extraction from twitter." (2012)

[14] Buscaldi, Davide, and HaïfaZargayouna. "Yasemir: Yet another semantic information retrieval system." Proceedings of the sixth international workshop on Exploiting semantic annotations in information retrieval. ACM, 2013.

[15] Mihalcea, Rada F., and Silvana I. Mihalcea. "Word semantics for information retrieval: moving one step closer to the Semantic Web." Tools with Artificial Intelligence, Proceedings of the 13th International Conference on. IEEE, 2001.

[16] EL-SALAM, Shimaa M. Abd, et al. Extracting Arabic Relations from the Web. arXiv preprint arXiv:1603.02488, 2016.