ISSN 1598-0170 (Print)
ISSN 2287-1136 (Online)
http://www.jksii.or.kr

NLP기반 NER을 이용해 소셜 네트워크의 조직 구조 탐색을 위한 협력 프레임 워크☆

A Collaborative Framework for Discovering the Organizational Structure of Social Networks Using NER Based on NLP

프랭크 엘리호데* 양 현 호** 이 재 완***
Frank I. Elijorde Hyunho Yang Jaewan Lee

요 약

방대한 양의 데이터로부터 정보추출의 정확도를 향상시키기 위한 많은 방법이 개발되어 왔다. 본 논문에서는NER(named entity recognition), 문장 추출, 스피치 태깅과 같은 여러 가지의 자연어 처리 작업을 통합하여 텍스트를 분석하였다. 데이터는 도메인에 특화된 데이터 추출 에이전트를 사용하여 웹에서 수집한 텍스트로 구성하였고, 위에서 언급한 자연어 처리 작업을 사용하여 비 구조화된 데이터로부터 정보를 추출하는 프레임 워크를 개발하였다. 조직 구조의 탐색을 위한 택스트 추출 및 분석 관점에서 연구의 성능을 시뮬레이션을 통해 분석하였으며, 시뮬레이션 결과, 정보추출에서 MUC 및 CoNLL과 같은 다른 NER 분석기 보다 성능이 우수함을 보였다.

ABSTRACT

Many methods had been developed to improve the accuracy of extracting information from a vast amount of data. This paper combined a number of natural language processing methods such as NER (named entity recognition), sentence extraction, and part of speech tagging to carry out text analysis. The data source is comprised of texts obtained from the web using a domain-specific data extraction agent. A framework for the extraction of information from unstructured data was developed using the aforementioned natural language processing methods. We simulated the performanceof our work in the extraction and analysis of texts for the detection of organizational structures. Simulation shows that our study outperformed other NER classifiers such as MUC and CoNLL on information extraction.

☞ keyword : Semantic Web (씨멘틱 웹), Social Network Analysis (소셜 네트웍 분석), Natural Language Processing (자연어 처리), Machine Learning (기계 학습)

1. 서 론

For a long time, information extraction has been the foundation of every knowledge discovery endeavors. The World Wide Web had already penetratedalmost every aspect of our society. Nowadays, social activities in the real world have been recorded in a massive number of inter-linked Web documents. Whether intentionally or unintentionally, these documents can provide information about what is going on, what has happened, as well as the people and things involved and their relationship with each other. When dealing with enormous data, we need to find a more efficient way to find the underlying relationships between concealed entities. Useful information can be derived from this enormous pool of data should there be a way of transforming it into a structured form that would provide adequate semantic metadata.

A number of methods have been developed to further improve the process of mining information from a vast amount of data sources. One of these methods is using text analysis, which in many studies has been proven to be an

* 정 회 원 : 군산대학교 전자정보공학부 박사과정
frank_elijorde@yahoo.com.ph
** 정 회 원 : 군산대학교 정보통신공학과 교수
hhyang@kunsan.ac.kr
*** 종신회원 : 군산대학교 정보통신공학과 교수
jwlee@kunsan.ac.kr (교신저자)

[2011/12/01 투고 - 2011/12/05 심사 - 2012/03/13 심사완료]

☆ This work (Grants No. 00047659) was supported by Business for Cooperative R & D between Industry, Academy, and Research Institute funded Korea Small and Medium Business Administration in 2011.

http://dx.doi.org/10.7472/jksii.2012.13.2.99

effective tool. Taking advantage of its capabilities, we further extended its use by applying with it a number of natural language processing tasks such as named entity recognition, sentence extraction, and part of speech tagging. From this standpoint, we were able to draw out results that could be of significant value to analysts. Looking at our results, combining the capabilities ofthe aforementioned processes would greatly improve the accuracy and reliability of text analysis in knowledge discovery. In this paper, we focused our methods in discovering organizational structures from a given corpus of text.

Social network describes agroup of social entities and the pattern of inter-relationships among them. What the relationship means varies, from those of social nature, such as kinship or friendship among people, to that of transactional nature, such as trading relationship between countries [1]. A considerable number of works has been done in the field of Social Network Analysis. However, existing network analysis tools used by law enforcement and intelligence agencies mainly focus on network visualization and do not have much structural analysis capability. Such a limitation might be successfully addressed by several methods from social network analysis research [2]. Therefore, an analyst's concrete understanding of the structural properties of a network would aid in the identification of valuable members to be subjected for removal or monitoring, as well as to apply disruptive measures to exposed vulnerabilities.

Texts are abundant sources of information about anything. Machine readable texts that convey information about covert networks are available on a large scale. In order to extract the organizational structure of covert networks effectively and efficiently from texts, appropriate tools and techniques are needed [3]. Over the web, a vast amount of text is available in electronic form that shows information about people, the groups in which they belong, the events or activities in which they are involved, time and place, as well as the resources in hand. Such data and its accessibility enable the development and evaluation of automated techniques for the efficient and effective extraction of the underlying social and organizational structures.

In this paper, we present a text-based approach to discovering organizational structures. A data extraction agent does the first step of the data gathering process. Unstructured data will be gathered in the form of texts from various sources over the web. The data pre-processing module takes part in the early processes of information extraction. Namely the tasks involved are Named Entity Recognition and Sentence Extraction based on NER. Finally, the data processing module performs the final stages. A process called Part-of-Speech Tagging is used to form statements between entities which will be later on classified as associations among entities.

2. Background and Related Work

2.1. Named Entity Recognition

Named Entity Recognition (NER) is a subtask of information extraction that seeks to locate and classify atomic elements in text into predefined categories such as the names of persons, organizations, and locations [4]. It was shown in [5] that focused named entities are useful for many natural language processing applications, such as document summarization, search result ranking, and entity detection and tracking. In [6], adaptive NER is used for social network analysis. The method utilized is link analysis based on clustering algorithm, which was used to find entities which are closely related to each other. Entities that occur in same documents are deemed to co-occur with each other. The software developed used relevant lexicons and patterns decided by the domain to perform NER. Although the approach was able to retrieve entities, it was largely dependent on domain-specific ontology, which is manually crafted and maintained.

2.2. Inter-worker Agreement

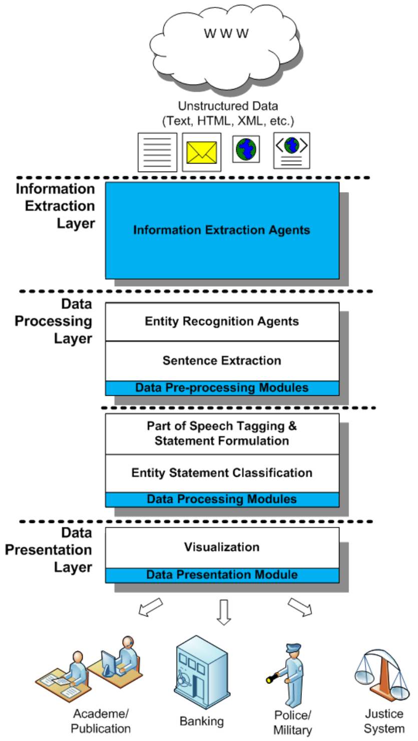
Another work in [7] with the goal of improving NER, made use of "mechanical turk". Workers are given data to work on, and from that their respectiveannotations will be evaluated. They utilized an algorithm to determine the quality of worker annotations. As a result, entities can be drawn out based on the annotation agreement among workers. However, this approach has posed a major issue: information loss. For instance, removing some existing model and inserting training data from some worker causes the model's performance to go down in predicting person entities, but increases its performance on organization entities.

2.3. Sentence Extraction and Part of Speech Tagging

A study in [8] shows that sentence extraction can capture the most salient pieces of information in the original document. In corpus linguistics, part-of-speech tagging, also called grammatical tagging or word-category disambiguation, is the process of marking up the words in a text as corresponding to a particular part of speech, based on both its definition, as well as its context [9]. Algorithms for the extraction of {subject, predicate, object}triplets from a given parse tree of a sentence was presented in [10]. A triplet is a representation of a subject-verb-object relation in a sentence, where the verb is the relation. A machine learning approach to extract subject-predicate-object triplets from English sentences were demonstrated in [11].

2.4. Text Analysis and Social Networks

A work in [12] focused in analyzing an online social visualization web site using word-based textual similarity measures to study the relationships among user comments. They also detect and visualize the patterns of user collaborations. Using word sense disambiguation the parts of speech and probable senses of the words in a comment are found. In [13], they proposed a method of social tension detection and intention recognition based on natural language analysis of social networks, forums, blogs and news comments. The approach combines natural language syntax and semantics analysis with statistical processing to identify possible indicators of social tension. A work in [14] adopted network analysis tools to carry out a terrorist social network quantitative analysis. Text information on terroristic activities from various network were processed by text data mining, and visual network models of the organizational structure of terrorists were made. The tools they utilized were ORA [15] and Automap [16].



(Figure 1) The proposed architecture.

2. NER Based on NLP

The proposed architecture is presented in a layered approach as shown in Figure 1. The main components of the architecture are distributed among the three layers, namely the Information Extraction Layer, Data Processing Layer, and Data Presentation Layer.

The Information Extraction Layer does the initial job of retrieving unstructured text resources from the web using an extraction agent. Each agent specializes on certain document formats found over the web, thus enabling it to "crawl"the web based on a given parameter. The Data Processing Layer is composed of two components, the Data Pre-processing Module and the Data Processing Module. The first task in the Data Pre-processing Layer is accomplished by the Named Entity Recognition (NER). It is utilized to identify the names of people, places, organizations, and other concepts, thereby establishing the foundation of the semantics. After which, sentences with relevant entities are extracted. Sentence Extraction is used to identify the most salient sentences of a text. It works as a filter which allows only important sentences to pass. In the Data Processing Module, the first task is Part of Speech Tagging (PoST). It is the process of marking up the words in a text as corresponding to a particular part of speech, based on both its definition, as well as its context. Entity Statement Classification is the second task in the Data Processing module. This process works by associating words identified as parts of speech. Finally, processed data is shown in a more intuitive form within the Data Presentation Layer through the Visualization Module.

3.1 Extracting Data from the World Wide Web

The information extraction agent is assigned to extract text on a given domain. In this work, we are interested in extracting documents that might contain information leading to the discovery of possible criminal organizations. Pages with interesting contents are being crawled, thus extracting text from it. The agents take unseen texts as input and produces fixed-format, unambiguous data as output. This data may be used directly for display to users, or may be stored in a database or spreadsheet for later analysis [17].

3.2 Identifying Entities and Extracting Relevant Sentences

As soon as the input file is ready, relevant entities are identified using NER. Entities are identified by the entity extraction agent based on a particular domain. This is a crucial process of data pre-processing since it will provide the basis for the succeeding tasks within the framework. The particular purpose of NER in this work is to identify entities such as person, location and organization. In defining the task, it is essential to recognize information units like names, including person, organization and location names, and numeric expressions including time, date, money and percent expressions [18].

After the previous process, sentence extraction is performed. In this routine, it is used to identify the most salient sentences of a text. Sentence extraction algorithms were originally used in the automatic summarization of documents which involves the creation of a shortened version of a text. In this approach, sentences with sufficient number of potential entities are drawn out. The product of this procedure contains the most important portions of the original text. In Figure 2, an algorithm designed for this specific purpose is shown.

```
let d=document
let s=sentence
let w=word
let e= ent_count
for each s in d
        for each w in s
            if NER(w) is entity then
                e++
        next w
        if e > 1 then
            s is relevant
            collect(s)

next s
```

(Figure 2) The algorithm for sentence extraction based on NER.

In figure 2, sentences in a document are read. Using the NER(w) function, a word is analyzed if it is an entity or not. A sentence which contains two or more entities is considered relevant and is therefore extracted.

3.3 Finding the Nodes and Establishing the Edges

The ability to recognize previously unknown entities is an essential part of NER systems. However, it is evident that even the most sophisticated NER systems are still susceptible to losses therefore affecting its accuracy [19]. These are the "non-coding"regions of NER. This applies to sequences where there was an entity and the system guessed it right, there was an entity but the system missed it, and also there was no entity but the system hypothesized one. These events are considered as labeling error, a boundary error, and a label-boundary error [20].

From the previous process, potential nodes have already been identified. To further enhance the entity extraction performed, we integrate Part of Speech Tagging into NER. The tagging process utilizes the Penn Treebank tag set for reading text and assigns parts of speech to each word, such as noun, verb, adjective, adverb, and determiner. Along the process, a number of tags are made. If we put it in the way humans analyze a sentence, we are only interested with parts of speech which would give us the subject, predicate and object of a statement. To realize this concept, we further filter out the product of PoST.

At this point, we are only interested on the words tagged as noun and verb, therefore words like adverbs, adjectives, and determiners were not considered. From this, we were left only with words tagged as noun and verb. The primary word orders that are of interest are then considered to form the subject, verb, and object of a sentence therefore forming the "triplet"of a sentence. The algorithm [10] in Figure 3 describes this process.

```
function TRIPLET-EXTRACTION(sentence) returns
a solution, or failure
result ←EXTRACT-SUBJECT(NP)
"ŧ EXTRACT-PREDICATE(VP)
"ŧ EXTRACT-OBJECT(NP)
if result ≠failure then return result
else return failure
```

(Figure 3) The algorithm for extracting triplets in treebank output.

In figure 3, the algorithm describes the process of forming the triplet of a sentence. The triplets of a sentence refer to the word order of subject, predicate, and object that presents an idea. The *EXTRACT-SUBJECT(NP)*function extracts the subject. It is the person or a thing that carries out the action. The *EXTRACT-OBJECT(VP)*function extracts the object. It is the person or a thing upon whom or upon which the action is carried out. The *EXTRACT-PREDICATE(NP)*function extracts the predicate. The predicate in a sentence tells about the action done to a person or to a thing.

3.4 Classifying Relationships Among Nodes

The final step of data processing is the formulation and classification of statements between entities. These statements are reflected in a matrix [3] that would contain entities as a set of rows and columns. This approach is best described by intersecting rows and columns, therefore allowing us to establish potential relationship among entities.

A statement can be classified as one involving a person, organization, resource, location, and some other entities. For example in figure 4, a person to person interaction is a possible social network, a person to organization interaction could tell us about membership to a network, a person to location interaction could lead us to a network's point of operation, and a person to resource interaction could let us identify resource providers.

3.5 Visualization of Organizational Structure

In the data presentation layer, the processed data is presented in a form that can be easily understood and analyzed. The visualization of relationships betweenentities

Entities	Person	Organization	Resources	Task	Event	Location
Person	X	X	X	X	X	X
Organization		X	X	X	X	X
Resources			X	X	X	X
Task				X	X	X
Event					X	X
Location						X

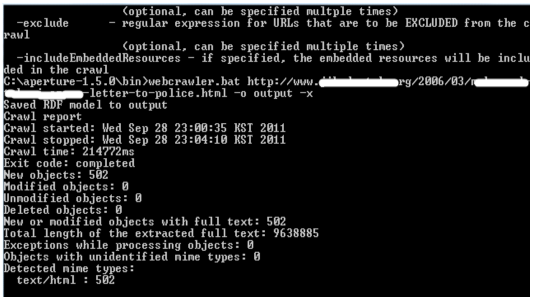
(Figure 4) Entity interaction formed by statements.

derived from the text is implemented by the data presentation module. Entities comprise the nodes of the network while the links between them are depicted as edges. The visualized information can now be used for further analysis and knowledge formation.

4. Implementation and Evaluation

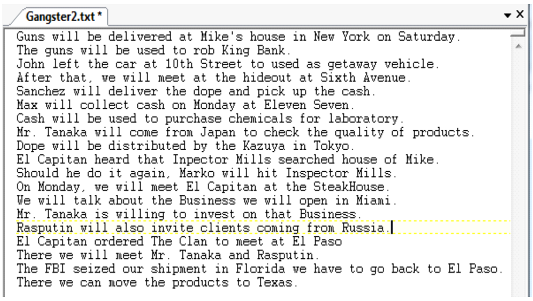
4.1. Implementation

This work was implemented using Java. Aperture [21] was used for the text extraction process of the data extraction layer. Aperture is a Java framework for extracting and querying full-text content and metadata from various information systems (e.g. file systems, web sites, mail boxes) and file formats (e.g. documents, images). The output of the "web crawling"process was written on a text file. The procedure in Figure 5 shows the extraction of text from a webpage. Extracted texts are then stored in a format ready for further processing. Preferably, text file is the best option for this purpose. Figure 6 shows an example of a text file to be processed.



```
                  (optional, can be specified multple times)
  -exclude      - regular expression for URLs that are to be EXCLUDED from the c
rawl
                  (optional, can be specified multiple times)
  -includeEmbeddedResources - if specified, the embedded resources will be inclu
ded in the crawl
C:\aperture-1.5.0\bin>webcrawler.bat http://www.[redacted]rg/2006/03/[redacted]
[redacted]-letter-to-police.html -o output -x
Saved RDF model to output
Crawl report
Crawl started: Wed Sep 28 23:00:35 KST 2011
Crawl stopped: Wed Sep 28 23:04:10 KST 2011
Crawl time: 214772ms
Exit code: completed
New objects: 502
Modified objects: 0
Unmodified objects: 0
Deleted objects: 0
New or modified objects with full text: 502
Total length of the extracted full text: 9638885
Exceptions while processing objects: 0
Objects with unidentified mime types: 0
Detected mime types:
  text/html : 502
```

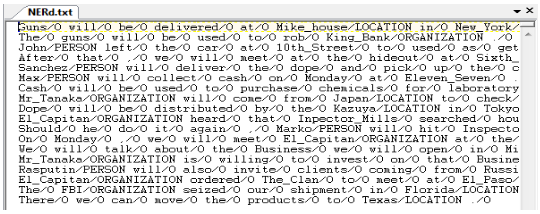
(Figure 5) Extracting text from a webpage.



```
Guns will be delivered at Mike's house in New York on Saturday.
The guns will be used to rob King Bank.
John left the car at 10th Street to used as getaway vehicle.
After that, we will meet at the hideout at Sixth Avenue.
Sanchez will deliver the dope and pick up the cash.
Max will collect cash on Monday at Eleven Seven.
Cash will be used to purchase chemicals for laboratory.
Mr. Tanaka will come from Japan to check the quality of products.
Dope will be distributed by the Kazuya in Tokyo.
El Capitan heard that Inpector Mills searched house of Mike.
Should he do it again, Marko will hit Inspector Mills.
On Monday, we will meet El Capitan at the SteakHouse.
We will talk about the Business we will open in Miami.
Mr. Tanaka is willing to invest on that Business.
Rasputin will also invite clients coming from Russia.
El Capitan ordered The Clan to meet at El Paso
There we will meet Mr. Tanaka and Rasputin.
The FBI seized our shipment in Florida we have to go back to El Paso.
There we can move the products to Texas.
```

(Figure 6) Text file to be used for processing.

The text file is then subjected to named entity recognition using Stanford NER [22]. Stanford NER is a Java implementation of a Named Entity Recognizer. It provides a general implementation of linear chain Conditional Random Field (CRF) sequence models, coupled with well-engineered feature extractors for Named Entity Recognition. Figure 7 shows the resulting text after NER was carried out.

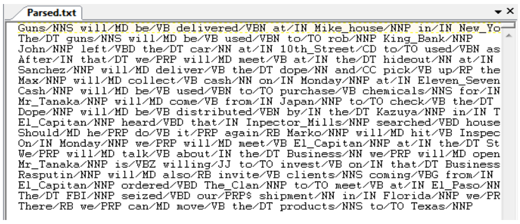


```
Guns/O will/O be/O delivered/O at/O Mike_house/LOCATION in/O New_York/
The/O guns/O will/O be/O used/O to/O rob/O King_Bank/ORGANIZATION ./O
John/PERSON left/O the/O car/O at/O 10th_Street/O to/O used/O as/O get
After/O that/O ,/O we/O will/O meet/O at/O the/O hideout/O at/O Sixth_
Sanchez/PERSON will/O deliver/O the/O dope/O and/O pick/O up/O the/O c
Max/PERSON will/O collect/O cash/O on/O Monday/O at/O Eleven_Seven/O .
Cash/O will/O be/O used/O to/O purchase/O chemicals/O for/O laboratory
Mr_Tanaka/ORGANIZATION will/O come/O from/O Japan/LOCATION to/O check/
Dope/O will/O be/O distributed/O by/O the/O Kazuya/LOCATION in/O Tokyo
El_Capitan/ORGANIZATION heard/O that/O Inpector_Mills/O searched/O hou
Should/O he/O do/O it/O again/O ,/O Marko/PERSON will/O hit/O Inspecto
On/O Monday/O ,/O we/O will/O meet/O El_Capitan/ORGANIZATION at/O the/
We/O will/O talk/O about/O the/O Business/O we/O will/O open/O in/O Mi
Mr_Tanaka/ORGANIZATION is/O willing/O to/O invest/O on/O that/O Busine
Rasputin/PERSON will/O also/O invite/O clients/O coming/O from/O Russi
El_Capitan/ORGANIZATION ordered/O The_Clan/O to/O meet/O at/O El_Paso/
The/O FBI/ORGANIZATION seized/O our/O shipment/O in/O Florida/LOCATION
There/O we/O can/O move/O the/O products/O to/O Texas/LOCATION ./O
```

(Figure 7) Text file applied with NER.

Part of Speech Tagging isthen used to mark words with their corresponding value in the Penn Treebank tag set. The PoST process was implemented using Stanford Log-linear Part-of-Speech Tagger [23]. It is a Java-made tagger that reads text and assigns parts of speech to each word and other token.

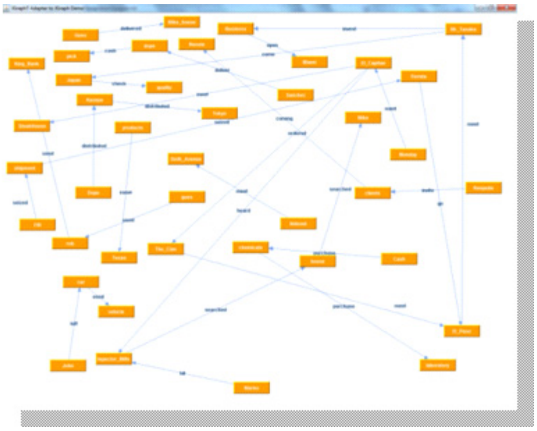
As shown in Figure 8, the tagging process utilized the Penn Treebank tag set for reading text and assigns parts of speech to each word, such as noun, verb, adjective, adverb, and determiner. From this process, triplets can be constructed out of the tagged words. For example, we have the sentence: "Mike bought a car in New York". Mike, car, and New York are nouns and bought is a verb. Further categorizing the tagged elements, Mike is a Person, car is a Resource, and New York is a Location. This will form an interesting interaction between entities, allowing us to come up with statements between them.



```
Guns/NNS will/MD be/VB delivered/VBN at/IN Mike_house/NNP in/IN New_Yo
The/DT guns/NNS will/MD be/VB used/VBN to/TO rob/NNP King_Bank/NNP
John/NNP left/VBD the/DT car/NN at/IN 10th_Street/CD to/TO used/VBN as
After/IN that/DT we/PRP will/MD meet/VB at/IN the/DT hideout/NN at/IN
Sanchez/NNP will/MD deliver/VB the/DT dope/NN and/CC pick/VB up/RP the
Max/NNP will/MD collect/VB cash/NN on/IN Monday/NNP at/IN Eleven_Seven
Cash/NNP will/MD be/VB used/VBN to/TO purchase/VB chemicals/NNS for/IN
Mr_Tanaka/NNP will/MD come/VB from/IN Japan/NNP to/TO check/VB the/DT
Dope/NNP will/MD be/VB distributed/VBN by/IN the/DT Kazuya/NNP in/IN T
El_Capitan/NNP heard/VBD that/IN Inpector_Mills/NNP searched/VBD house
Should/MD he/PRP do/VB it/PRP again/RB Marko/NNP will/MD hit/VB Inspec
On/IN Monday/NNP we/PRP will/MD meet/VB El_Capitan/NNP at/IN the/DT St
We/PRP will/MD talk/VB about/IN the/DT Business/NN we/PRP will/MD open
Mr_Tanaka/NNP is/VBZ willing/JJ to/TO invest/VB on/IN that/DT Business
Rasputin/NNP will/MD also/RB invite/VB clients/NNS coming/VBG from/IN
El_Capitan/NNP ordered/VBD The_Clan/NNP to/TO meet/VB at/IN El_Paso/NN
The/DT FBI/NNP seized/VBD our/PRP$ shipment/NN in/IN Florida/NNP we/PR
There/RB we/PRP can/MD move/VB the/DT products/NNS to/TO Texas/NNP
```

(Figure 8) The text file from 3.2 applied with PoST.

The statements formed between entities are classified into relationships. A triplet like(Mike, bought, car) can form a relationship. This can be considered as an interaction between a person and a resource. Another example, "Rasputin will also invite clients coming from Russia."The triplet (Rasputin, invite, clients) is a potential relationship. This can be classified as an interaction between persons, therefore revealing a possible social network.



(Figure 9) Organizational Structure.

The visualization of the network structure was implemented using JGraph. It is a powerful, lightweight, feature-rich, and thoroughly documented open-source graph component available for Java [24].

The entities identified in the analysis of data were depicted as nodes, and on the other hand, the edges represent the links between entities. Figure 9 shows the visualization of a possible organizational structure.

4.2. Evaluation

In the actual run of the system, a sample document was fed. The domain of the document is focused on monitoring the activities of a suspected mafia. The actual count of entities was setto 41 which will be used as a basis for evaluating the actual performance of the system. For the purpose of evaluation, the text was fed onto NER using three different classifier models included with Stanford NER. The performance is then measured in terms of Precision, Recall, and F-measure as shown by the following equations:

$$precision = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

$$recall = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

$$F = \frac{2 \ * \text{precision} \ * \text{recall}}{\text{precision} + \text{recall}}$$

The three NER classifiers were tested and their performance was measured. The first model usedis a seven-class model trained for MUC. It has seven classes- Time, Location, Organization, Person, Money, Percent, and Date. Using this model, the NER routine was able to retrieve a number of entities. In this case, it was able to retrieve 21 entities and 5 of these were mislabeled. The second model used is a four-class model trained for CoNLL. It has four classes- Location, Person, Organization, and Misc. Using the said model, the NER routine was able to retrieve more entities, this time it was able to retrieve 25 entities but 6 were mislabeled. The third model used as classifier is a combination of the previous models; it has three classes- Location, Person, and Organization. Using the said model, the NER routine was still able to retrieve 19 entities but 2 of them were mislabeled.

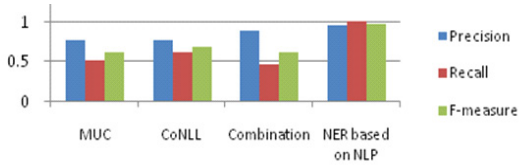
Finally, entity recognition using NER based on NLP was tested. It was able to tag 43 entities, however; there was an excess of 2 which should have not been included. Despite of this, still it remarkably improved the process of extracting entities based on the following results:

(Table 1) Simulation results

Algorithm	Precision	Recall	F-measure
MUC	0.761905	0.512195	0.61258
CoNLL	0.76	0.609756	0.676638
Combination	0.894737	0.463415	0.610586
NER based on NLP	0.95122	1.00	0.975

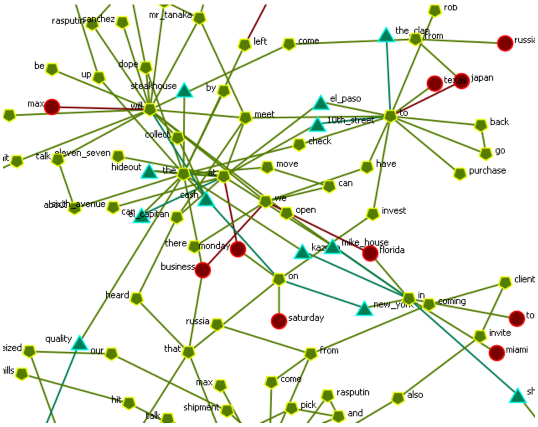
In table 1, it is shown that in the first three methods

CoNLL has the highest f-measure. This is due to the balance between its precision and recall. On the other hand, MUC has a slightly higher f-measure as compared to Combination. However, compared to MUC and CoNLL, Combination has a higher precision. Despite of this, it suffered in the area of its recall which significantly affects its accuracy. Using NER based on NLP, the highest precision and recall was achieved with values of 0.95122 and 1.00 respectively. It is actually trivial to achieve 100% recall by returning all relevant documents in response to a query. This means that recall alone is not enough so we need to measure the number of non-relevant documents also, for example by computing the precision. Once the precision and recall are known, the f-measure is then derived. In figure 10, a graphical representation of the values is shown.



(Figure 10) Summary of performance evaluation

We also tested the tools utilized in [14] for text analysis and social network visualization. The method was able to retrieve 15 entities with a Precision, Recall, and F-measure of 1.0, 0.37, and 0.54 respectively. A network visualization using their tool was also generated as shown in figure 11.



(Figure 11) Visualization generated by ORA.

In its own right, the tool was indeed powerful. However it is only applicable for entity pairs which will comprise the node links. On the other hand, our work makes use of word triplets consisted of two entity names and a verb. Using triplets, we can visualize not only the link but also the event that occurred between the two entities.

4. Conclusions and Future Work

Studies show that even the most advanced NER systems are brittle. It means that NER systems developed for a particular domain do not typically perform well on other domains. This simply shows that NER systems are always dependent on how "well-trained"its classifier is, which is only limited to a particular domain in which it is directed to learn.

In this work, part of speech tagging has a significant impact on entity recognition in the sense that the information missed by NER can be retrieved by PoST without having to depend on a domain-specific entity classifier. This can be useful in extracting information from unstructured documents in which unknown entities can be revealed and analyzed.

The significant contribution of this work is the improvement of the accuracy of information extraction by using natural language processing techniques in processing data. As shown by the results, accuracy of text analysis is remarkably enhanced. To further assist with the analysis, a visualization of the structure formed by the links and entities was also provided.

The efficiency and accuracy of the proposed text analysis technique applied on unstructured data sources could provide indispensable assistance to crime analysts. This enables the extraction of hidden crime-related information from unlikely sources as well as to generate and visualize their potential networks. Our work can also be used to monitor and analyze suspicious Web sites as well as threatening online social activities. Practically, it would provide immediate discovery of knowledge pertaining to criminal organizations that would take a lot of time and effort when done manually.

In our future work, we intend to expand the proposed framework by integrating a knowledge broker in the data presentation layer. This would provide a domain expert that would facilitate the extraction of knowledge from areas specified by its clients. With such capability, mining of significant information over an enormous and unstructured data source will be a straightforward task.

References

[1] H. Lauw, E. Lim, T. Tan, and H. Pang: Mining Social Network from Spatio-Temporal Events, Proceedings of SIAM Data Mining Conference (2005)

[2] J.J. Xu and H. Chen:Crimenet Explorer: A Framework For Criminal Network Knowledge Discovery., ACM Transactions on Information Systems, pp. 201–226 (2005)

[3] J. Diesner, and K.M. Carley: Using Network Text Analysis to Detect The Organizational Structure of Covert Networks, Proceedings of the North American Association for Computational Social and Organizational Science (NAACSOS) Conference, Pittsburgh, PA (2004).

[4] Named Entity Recognition, http://en.wikipedia.org/wiki/Named_entity_recognition

[5] L. Zhang, Y. Pan, and T. Zhang:Focused Named Entity Recognition using Machine Learning, SIGIR’04 (2004)

[6] J. Zhu, A. L. Goncalves, and V. Uren: Adaptive Named Entity Recognition for Social Network Analysis and Domain Ontology Maintenance, Tech Report kmi-04-30 (2005)

[7] W. Murnane:Improving Accuracy of Named Entity Recognition on Social Media Data, Thesis, Graduate School, University of Maryland (2010)

[8] K. Knight, and D. Marcu:Summarization beyond sentence extraction: A probabilistic approach to sentence compression, Artificial Intelligence Volume 139, Issue 1, pp. 91-107 (2002) 8

[9] Part of Speech Tagging, http://en.wikipedia.org/wiki/Part-of-speech_tagging

[10] D. Rusu, L. Dali, B. Fortuna, M. Grobelnik, and D. Mladenid: Triplet Extraction from Sentences, In Proceedings of the 10th International Multiconference "Information Society--IS 2007". Vol. A, pp. 218-222 (2007)

[11] L. Dali and B. Fortuna: Triplet extraction from sentences using svm. In SiKDD (2008)

[12] . Karmakar, and Z. Ying, "Mining collaboration through textual semantic interpretation,"Intelligent Systems (HIS), 2011 11th International Conference onvol., no., pp.728-733, 5-8 Dec. 2011

[13] O. Vybornova, I. Smirnov, I. Sochenkov, A. Kiselyov, I. Tikhomirov, N. Chudova, Y. Kuznetsova, G. Osipov, "Social Tension Detection and Intention Recognition Using Natural Language Semantic Analysis: On the Material of Russian-Speaking Social Networks and Web Forums,"and Security Informatics Conference (EISIC), 2011 Europeanvol., no., pp.277-281, 12-14 Sept. 2011

[14] Sun Duo-Yong; Guo Shu-Quan; Zhang Hai; Li Ben-Xian; , "Study on covert networks of terroristic organizations based on text analysis,"Intelligence and Security Informatics (ISI), 2011 IEEE International Conference onvol., no., pp.373-378, 10-12 July 2011

[15] Automap by CASOS, http://www.casos.cs.cmu.edu/projects/automap/

[16] ORA by CASOS, http://www.casos.cs.cmu.edu/projects/ora/

[17] H. Cunningham: Information Extraction-A User Guide, Research memo CS–97–02 (1997)

[18] D. Nadeau, and S.Sekine: A survey of named entity recognition and classification, Lingvisticae Investigationes, Volume 30,1 , pp. 3-26(24) (2007)

[19] D. Nadeau, and S.Sekine: A survey of named entity recognition and classification, Lingvisticae Investigationes, Volume 30,1 , pp. 3-26(24) (2007)

[20] Doing Named Entity Recognition? Don’t optimize for F1, http://nlpers.blogspot.com/2006/08/doing-named-entity-recognition-dont.html

[21] Aperture Framework, http://aperture.sourceforge.net/

[22] Stanford Named Entity Recognizer, http://nlp.stanford.edu/software/CRF-NER.shtml

[23] Stanford Log-linear Part-Of-Speech Tagger, http://nlp.stanford.edu/software/tagger.shtml

[24] Graph, http://sourceforge.net/projects/jgraph

◐ 저 자 소 개 ◑

Frank I. Elijorde

2003년 Western Visayas College of Science and Technology, Philippines, BS in Information Technology
2007년 Western Visayas College of Science and Technology, Philippines, MS in Computer Science
2011~현재 Kunsan National University, South Korea, Graduate Student in Ph. D. Course
관심분야 : Distributed systems, data mining, social networks, ubiquitous sensor networks, RFID.
E-mail : frank_elijorde@yahoo.com.ph

양 현 호 (Hyunho Yang)

1986년 광운대학교 전자공학과 졸업(학사)
1990년 광운대학교 대학원 전자공학과 졸업(석사)
2003년 광주과학기술원 정보통신공학과 졸업(박사)
1989~1990 삼성SDS 근무
1991~1997 포스데이타(주) 근무
1997~2005 순천청암대학 근무
2005~현재 군산대학교 정보통신공학과 교수
관심분야 : 무선데이터통신, RFID/USN etc.
E-mail : hhyang@kunsan..ac.kr

이 재 완 (Jaewan Lee)

1984년 중앙대학교 이학사-전자계산학
1987년 중앙대학교 이학석사-전자계산학
1992년 중앙대학교 공학박사-전자계산학
1996년 3월~1998년 1월 한국학술진흥재단 전문위원
1992년~현재 군산대학교 교수
관심분야 : 분산 시스템, 운영체제, 실시간 시스템, 컴퓨터 네트워크, 클라우드 컴퓨팅 등
E-mail: jwlee@kunsan.ac.kr