SVM based Extraction of Spatial Relations in Text

Xueying Zhang[1], Chunju Zhang[2], Chaoli Du[3], Shaonan Zhu[4]

Key Laboratory of Virtual Geography Environment, Nanjing Normal University, Nanjing 210046, China

[1]zhangsnowy@163.com

[2]zcjtwz@sina.com, Corresponding author

[3]viekiedu@gmail.com

[4] zhushaonan@gamil.com

***Abstract*—Natural language text describes the nature of people's internal representation of space. It is investigated that 80% of unstructured text has location expressions e.g. place names and spatial relations. In the past few years, text has become a most important geospatial data resource as well as survey, map, satellite images and GPS. The most previous research focused on the recognition of place names in text and its integration with map services. Spatial relations play an important role in the fields of spatial data modelling, spatial query, spatial analysis, spatial reasoning and map generalization. Spatial relations in text are described in natural language with qualitative spatial expressions including place names, spatial terms, prepositions, verbs and so on. And these expressions are combined with certain syntactic patterns to represent their semantic functions. An instance of spatial relation in text can be simply formalized as (C1, P1, C2, P2, C3), where P1 and P2 are place names, and C1, C2 and C3 are the context. Support Vector Machine (SVM) is a pattern recognition method popularly used in information extraction from text. This paper investigates the extraction of spatial relations based on SVM model which can implement the recognition of spatial expressions and their classification synchronously. For the SVM model, a set of feature vectors are specified, such as lexical tokens, spatial terms, syntactic structures and geographical feature types of place names, and a multi-label classifier is presented to solve the multi-classification problem. Finally, an experimental evaluation is explored in a Chinese annotation corpus. This study proves that spatial terms are important indicators for identification of spatial relations in text. However, there is serious ambiguity of their classification. Therefore, integration of much more context information could potentially improve the performance of extraction of spatial relations in text.**

***Keywords*—Spatial relation; spatial terms; text; support vector machine; feature vector**

I. INTRODUCTION

Text describes the nature of people's internal representation of space in natural language. The primary means for exchanging spatial information is natural language text. It is investigated that 80% of unstructured text has location expressions，such as place names, spatial relations and geographical attributes [1]. Especially, with the dramatic increase of World Wide Web and the fast development of Location-Based-Service (LBS), there have been growing spatial information in large text documents among people's daily life, such as yellow pages, path and address expressions. Text has become a most important geospatial data resource as well as survey, map, satellite images and GPS.

Location expressions mainly include specific place names, spatial relations and geographical attributes. People get used to describe place names with relative location information, especially spatial relations. Spatial relations play an important role in spatial data modelling, spatial query, spatial analysis, spatial reasoning and map generalization [2] The previous contributions mainly focused on the recognition of place names in text and its integration with map services [3, 4]. Therefore, it is significant to interpret spatial relations in text to improve the intelligent location-based services, geographical information retrieval, spatial natural language query and spatial scene reconstruction.

This paper proposes a supervised machine learning approach of Support Vector Machine (SVM) for extraction of spatial relations from text. The remainder of this paper is organized as follows. Section 2 surveys the previous work on information extraction and relation extractions. In section 3, SVM-based extraction of spatial relations is discussed which contains data pre-processing, SVM models and feature vectors. We conduct a performance evaluation of the proposed approach and error analysis on a Chinese annotation corpus in section 4. Finally, in section 5, a conclusion and discussion of future work is followed.

II. RELATED RESEARCH

Relation extraction has been a great concern in information extraction (IE) and was formulated as part of Message Understanding Conferences (MUC). It typically includes physical relations, personal or social relations, and membership between two entities in natural language. Work on relation extractions over the last two decades has progressed from linguistically unsophisticated models to the adaptation of natural language processing (NLP) techniques that use shallow parsers or full parsers and complicated machine learning methods [7]. There have been several typical models for relation extractions, such as rule-based models [5], feature vector-based models [6], SVM-based models [7] and kernel-based models [8]. However, most of these models focus on a single-class classification.

Spatial relations are the associations or connections between different real-world features, mainly including direction relations, topology relations and distance relations [9]. It is a fundamental issue for spatial data organization, retrieval, analysis and reasoning. The previous research has made much effort on computational models of spatial terms

978-1-4244-8351-8/11/$26.00 ©2011 IEEE

[10], spatial relation models [11], natural language spatial relation description [12], and semantic expressions of spatial relations [13]. Only a few documents discussed rule-based approach to extract spatial relations from text, which depends on manually induction of spatial terms and syntactic patterns of the spatial relation description [13, 14]. Obviously, these approaches are time-consuming and usually achieved unsatisfactory performance.

In text, spatial relations are described in a flexible way because of spatial cognition and linguistic customs. In most cases, a direction term indicates both implicit topology relations and explicit direction relations. Therefore, one spatial relation instance may correspond to several spatial relation categories. For example, "青龙镇位于上海市西部青浦县东北(The Qinglong Town is located in the northeastern border of Qingpu County which is in the western of Shanghai City)". There are three spatial relations between Qinglong Town and Qingpu County, i.e. the northeast direction relation, disconnected and externally connected topological relations. This is a multi–label problem which is more difficult than traditional single-label phenomenon in NLP and IE. To solve this problem, this paper presents a SVM-based model for extraction of spatial relations (i.e. recognition and classification) synchronously. A set of feature vectors are specified, e.g. lexical tokens, spatial terms, syntactic structures and geographical feature types of place names. Especially, a multi-label classifier is presented in order that spatial knowledge and linguistic knowledge could be introduced to improve the performance of the proposed model.

III. SVM-BASED EXTRACTION OF SPATIAL RELATIONS

SVM is an optimal classifier with a maximal margin in feature space. The extraction of spatial relations from NLP tasks typically represents instances by very high dimensional but very sparse feature vectors [15]. So SVM is suitable for spatial relation extraction from text. Extraction tasks are divided into two individual subtasks: spatial relation recognition and spatial relation classification. Spatial relation recognition is involved in identifying spatial relations from every pair of place names, especially, for the one which can fall into more than one spatial relation categories. The classification task aims to assign a specific category to each detected spatial relation instance. For each task, distinct features are applied ranging from lexical tokens to syntactic structures as well as spatial terms. A multi-label classifier is presented based on the traditional binary classifier of SVM to solve the multi-class classification. In this section, we first provide the overview of the framework of the extraction of spatial relations (Fig. 1). Then, we explain the details for each module in the framework.

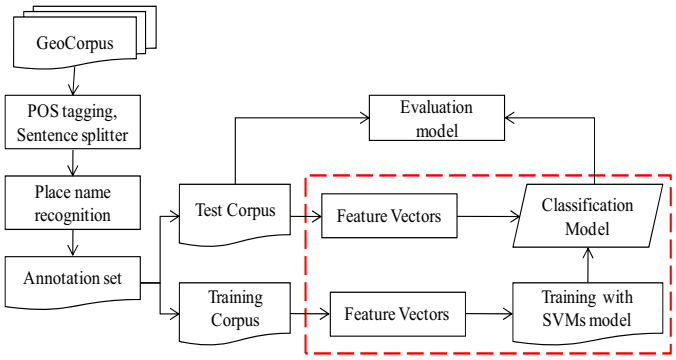

Fig.1 Framework of SVM-based extraction of spatial relations

A. SVM Model

Instances of spatial relations in text could be represented as (C1, P1, C2, P2, C3), where P1 and P2 are place names, and C1, C2 and C3 are the context. Supposing that $X = \{x_i\}_{i=1}^{n}$ is a dataset of all place names, in which x_i represents the first *i* pair of place names in the context, *n* is a total number of the pair of place names. The first *m* instances in the dataset are labelled with y_g ($y_g \in \{r_j\}_{j=1}^{R}$), in which r_j represents spatial relation categories and *R* is their total number. There are *u* (*u = n-m*) instances which are not labelled in the *X* dataset. Therefore, the extraction of spatial relations is to detect and classify the proper type of spatial relations for the *u* instances of spatial relations by learning the *m* labelled instances of spatial relations.

The SVM model is a pattern recognition method based on the principle of structural risk minimization. It maps the vector to a higher dimensional space, and establishes a maximum interval hyperplane in this space [16]. For a dataset of non-linear relationship, SVM will firstly map an input vector to a high-dimensional feature space by linear transformation. Then the input vector will obtain a linear relationship and a nonlinear classification. In formula (1) and (2), Φ is a non-linear relationship in a 2-dimensional space, after SVM processing it is transformed to a linear relationship in a 3-dimensional space.

$$\Phi: \ R^2 \to R^3 \tag{1}$$

$$(x_1, x_2) \to (z_1, z_2, z_3) := (x_1^2, \ \sqrt{2}x_1x_2, \ x_2^2) \tag{2}$$

In general, mapping a sample space of R^n into a high dimensional feature space of *H* could be denoted as $\Phi: \ R^n \to H$, in which *Φ* represents the mapping process. When mapping samples in a high dimensional space of *H*, SVM uses an inner product form of $< \Phi(x_i), \ \Phi(x_j) >$ during both training and testing process. Also, if there is a mapping of $k(x, x') = \langle \Phi(x), \ \Phi(x') \rangle$, then $x \to \Phi(x)$ could be indirectly mapped by a kernel function of $k(x, x')$. Linear features could be obtained in the high dimensional feature space.

One spatial relation may belong to several different categories in natural language expressions. Figure 2 shows an example of "青龙镇位于上海市西部青浦县东北境 (The

Qinglong Town is located in the north-eastern border of Qingpu County which is in the western of Shanghai City)".

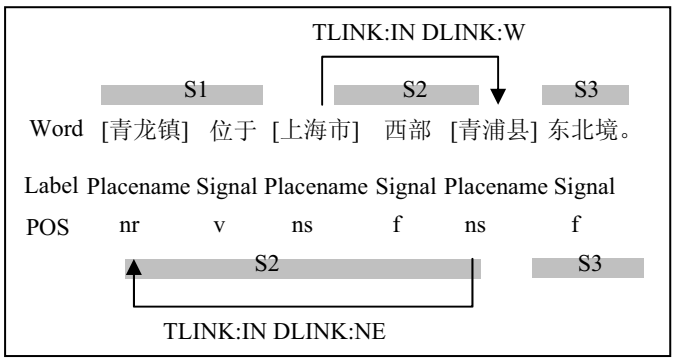

Fig.2 A spatial relation instance with different categories

In this example, there are two instances of spatial relations, one is <青浦县(Qingpu County), 青龙镇(Qinglong Town)> and the other is <上海市(Shanhai City), 青浦县(Qingpu County)>. The categories of spatial relations of <青浦县(Qingpu County), 青龙镇(Qinglong Town)> are a topological relation(TLINK) of tangential and non-tangential proper parts (IN) and a direction relation (DLINK) of northeast(NE). Spatial relations of < 上海市(Shanhai City), 青浦县 (Qingpu County)> are a topological relation(TLINK) of tangential and non-tangential proper parts (IN) and a direction relation (DLINK) of west (W). S1, S2, S3 are the context before and after place names while the label of "Signal" represents spatial terms.

Therefore, different from traditional single-label classifications, this phenomenon is a complex multi-label classification. SVM is a machine learning method of a binary classification essentially. To extract multi-class spatial relations, the following is a method for extending the single-label classification to a multi-label classification.

It is assumed that L represents a spatial relation dataset. If binary classifiers with a number of $|L|$ are to be learned, then each different l in L dataset will corresponds to a binary classifier of $H_1 : X \to \{l, \neg l\}$ separately. These L dataset need to be transformed to a new training data Dl before training. *The* total number of Dl is $|L|$. Each Dl dataset contains all instances in the initial training dataset L. If each instance in L is with the label l, then it is also with the label l in Dl dataset, else with a label $\neg l$ in Dl dataset. Therefore, the traditional binary classifier can be trained for each Dl dataset. For a new instance x, the training output will be a union set of the classification results from each training binary classifier $|L|$ (Formula 3).

$$H(x) = \bigcup_{l \in L} \{l\} : \; H_l(x) = l \tag{3}$$

B. Data Pre-processing

In this paper, we choose "Chinese Encyclopaedia (Geography)" called GeoCorpus as the training and test dataset, which includes 188 Chinese documents (91504 Chinese characters). The pre-processing includes the following steps:

- ICTLCAS platform of natural language processing is used to pre-process the original documents, e.g. word segmentation and parts of speech. The ICTLCAS is developed by the Institute of Computing Technology of Chinese Science Academy.
- Place names are recognized with the conditional random field (CRF) model beforehand. So all information regarding place names are available when spatial relations are extracted [4]. Then place names are annotated with a tag of <GNE> in brackets.
- Sentence boundaries should be detected. One spatial relation normally involves a pair of place names and a few spatial terms, and these elements are described within one sentence. The detection of sentence boundary could be performed by some simple rules, such as the punctuation of '?', '!', or '.'. At this step, a sentence with less than two place names will be filtered.
- Categories of Spatial relations and spatial terms are identified as a knowledge base. There are more than 20 categories of spatial relations [15]. For example, topology relations are divided into extended connection (EC), discrete connection (DC), partially overlap (PO), equality (EQ), etc, and direction relations are divided into southwest (SW), northwest (NW), east (E), west (W), etc. Each category of spatial relation includes more than one spatial term, and each spatial term may belong to more than one spatial relation category.
- The GeoCorpus data is divided into two parts, one is a training set and the other is a test set.

C. Feature Vectors

Feature vector is a numerical representation of instances. It means that an instance is converted to a feature vector x of n-dimension, in which x_i is the first i elements. Machine learning method based on feature vectors is to learn a classifier f from training datasets of (x_1, y_1), (x_2, y_2), $\cdots$, (x_n, y_n), then f can classify a new feature vector properly. Ideally, the more features we can provide, it would be more helpful for modelling extraction tasks of spatial relations. A majority of textual spatial relations are described with multi-qualitative spatial expressions, including linguistic units of place names, spatial terms, prepositions, verbs and so on. These units are combined with certain syntactic patterns to represent their semantic functions. The semantic function of a single spatial relation can be simply formalized as a quadruple, i.e. place name A, spatial terms, place name B, spatial reference. Therefore, the extraction of spatial relations can be transformed to identify place names, spatial terms and their semantic relations. By learning the semantic knowledge of spatial terms and instances of spatial relations a multi-label classifier is presented, and the feature sets are as following:

1) *Part of Speech (POS):* POS of both place names as well as all words between them are features. Although taking POS as a feature is very sparse as a whole, however, orientation words, prepositions and verbs can reflect spatial relations preferably.
2) *Place Names:* Lexical token of both place names as well

as all the words between them are features. If two or more words constitute a place name, each individual word is a separate feature.

3) *Mention Type of Place Names:* A place name mentioned can be named, nominal, or pronominal.
4) *Geographical Feature Categories:* There are about 40 categories of place names, such as river, lake, ocean, resident, attraction, traffic, organization, religious facility, etc.
5) *Categories of Spatial Relations:* There are more than 20 categories of spatial relations, such as EC, DC, IN, PO, N, NW, S, SW, SE.
6) *Spatial Terms:* Spatial terms have a strong indication for the expression of spatial relations, and are specified as feature vectors. If two or more words constitute spatial terms, each individual word is a separate feature.

Based on above features, feature vectors in the example of "青龙镇位于上海市西部青浦县东北境(The Qinglong Town is located in the northeastern border of Qingpu County which is in the western of Shanghai City)" are as following:

_POS_S1_NA_POS_S2_v/ns/f_POS_S3_f
_POSPLACE_S1_NA_POSPLACE_S2_v/place/f_POSPLACE_
S3_f_PLACE_青龙镇_PLACE__青浦县
_PLACEType_National Administrative Region PLACEType_
National Administrative Region
_SpatialTerms_S1_NA_SpatialTerms_S2_位于/西部
_S3_SpatialTerms_东北境
_SpaitalTerms_S1_linktype_NA_SpatialTerms_S1_direction_NA
_SpatialTerms_S2_linktype_IN/IN|DC _ SpatialTerms_ S2_
direction _NA/W
_SpatialTerms_S3_linktype_EC_SpatialTerms_S3_direction_NE

In the example, S1, S2, S3 are the context before and after place names, "National Administrative Region" represents the geographical feature category of place names, NA is an empty feature vector.

IV. Experimental Evaluation

In the experiment we evaluate extraction performance by a quantitative comparison of manual annotation results and automatic extraction results by the SVM model. The performance evaluations are defined as hamming loss, precision, recall, and F-measure (Formula 4-7). It is supposed that $|D|$ is a multi-label dataset. Multi-label instances with a number of $|D|$ can be expressed as $(x_i, Y_i), i = 1..|D|, Y_i \subseteq L$, in which L is a dataset with different labels. Supposing that H is a multi-label classification, and Z_i=H (x_i) is an annotation dataset of the instance x_i predicted by H classifier. In formula 4, $\triangle$ represents a symmetric difference of two datasets which corresponds to the boolean logic operation of *XOR*.

$$Hamming\ Loss(H,D) = \frac{1}{|D|}\sum_{1}^{|D|}\frac{|Y_i \triangle Z_i|}{|L|} \quad (4)$$

$$Precision(H,D) = \frac{1}{|D|}\sum_{1}^{|D|}\frac{|Y_i \cap Z_i|}{|Z_i|} \quad (5)$$

$$Recall(H,D) = \frac{1}{|D|}\sum_{1}^{|D|}\frac{|Y_i \cap Z_i|}{|Y_i|} \quad (6)$$

$$F-measure = \frac{2 \times Precision \times Recall}{(Precision + Recall)} \quad (7)$$

In order to focus on the performance evaluation of the extraction of spatial relations, it is supposed that all place names in the text have been recognized without mistakes. This paper takes 3/4 documents of the GeoCorpus as a test dataset and the remaining 1/4 as test data randomly. The training dataset contains a total number of 2515 instances of spatial relations, while test dataset contains 813 spatial relations. A 4 fold cross-validation is implemented in this experiment. The extraction results of spatial relations are with the hamming loss of 0.119, recall of 0.678, precision of 0.652 and F-measure of 0.661.

Following is an example of extraction of spatial relations. The highlight sentences are annotated the spatial terms and geographical entities. Figure 3 shows spatial relations extracted by the SVM model.

庐山是中国东南部江西省北部名山，位于九江县以南，星子县以西，风景区总面积 302 平方公里，山体面积 282 平方公里。庐山最高峰汉阳峰海拔 1474 米，东偎鄱阳湖，南靠南昌滕王阁，西邻京九大通脉，北枕滔滔长江。耸峙于长江中下游平原与鄱阳湖畔。(*Lushan mountain is a famous mountain in the north of Jiangxi province which located in the southeast part of China, located in the south of Jiujiang county, in the west of Xingzi county, the total area of its scenic is 302 square kilometers, the mountain area is 282 square kilometers. The highest peak of Lushan mountain is Hanyang Peak with a sea level of 1474 meters, cuddles the Poyang Lake in the east, leans against the Nanchang Poetic Pavilion in the south, adjacent to the Beijing-Kowloon big Tongmai in the west, surgings the Yangtze River in the north. Standingon the banks of the Yangtze River and the plain of Poyang Lake.)*

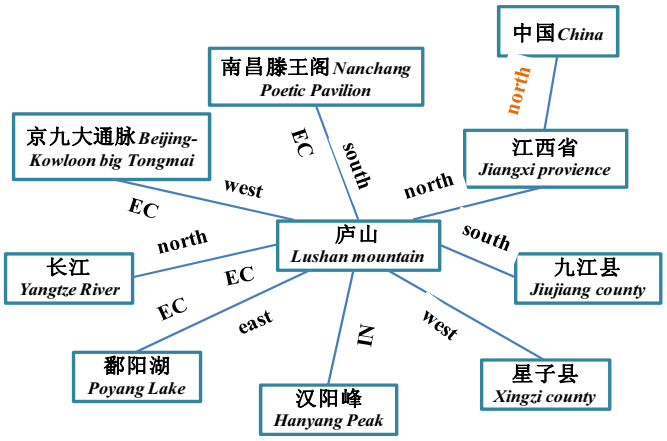

Fig. 3 Spatial relations extracted from the example

This experiment results indicate that the proposed SVM model can effectively extract spatial relations from text. However, spatial relations in natural language text are flexible. For example, some spatial relations are crossing sentences and with omitted place names, and even with complex route descriptions, etc. All of the spatial relations with this

phenomenon could not be identified effectively. Moreover, machine learning methods of the SVM model highly depends on the scale and quality of training datasets. As in figure 3, the identification of spatial relation between 江西省（Jiang Xi province）和中国(China) is wrong, and spatial relations in the sentence of "耸峙于长江中下游平原与鄱阳湖畔 (Standing up there on the banks of the Yangtze River and the plain of Poyang Lake)" are not extracted because Lushan mountain is not explicitly described in this sentence.

V. CONCLUSION

Interpretation of spatial relations is meaningful for intelligent location-base service, geographical information retrieval, spatial natural language query and supplement of spatial data. This paper presents a SVM-based model to extract spatial relations from text. The experimental evaluation is explored in a Chinese annotation corpus. Spatial relation expressions in natural language text are flexible. This study indicates that machine learning models such as SVM could effectively extract spatial relations in text. However, its performance greatly depends on the coverage of spatial terms and the accuracy of annotated corpus.

ACKNOWLEDGMENTS

This work was supported in part by the China National High-tech R&D Program under grant number 2007AA12Z221 and the National Nature Science Foundation under grant numbers 40971231. The authors wish to thank Dr. Junsheng Zhou for his insightful comments and suggestions.

REFERENCE

[1] B. Palkowsky and I. MetaCarta. A New Approach to Information Discovery—Geography Really Does Matter. *In Proceedings of the SPE Annual Technical Conference and Exhibition*, 2005.

[2] J. Chen and R. L. Zhao. Spatial Relations in GIS: A Survey on its Key Issues and Research Progress. *Acta Geodaetica et Cartograohica Snica*, 1999, Vo l. 28, No. 2, pp.96-102.

[3] Y. S. Li, Z. W. Yuan, X. Y. Zhang. Study on Geographical Entity Recognition in GIS. *Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition)*, 2008, Vol.20, No.6, pp. 719-724.

[4] X. Y. Zhang, G. N. Lv, Z. R. Xie. Extraction and Visualization of Geographical Names in Text. *In Proceedings of the 24th International Cartography Conference*, 2009, November 15-21, Santiago, Chile.

[5] S. Miller, H. Fox, L. Ramshaw, etc. A Novel Use of Statistical Parsing to Extract Information from Text. *In Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2000, pp. 226–233.

[6] G. D. Zhou, J. Su, J. Zhang, etc. Exploring Various Knowledge in Relation Extraction. *In Proceedings of ACL,2005*, USA, 2005, pp.427–434.

[7] G. Hong. Relation Extraction Using Support Vector Machine. *Berlin：Springer Berllin/Heidelberg*, 2005, pp. 366 – 377.

[8] R. Bunescu and R. Mooney. Subsequence Kernels for Relation Extraction. *In Proceedings of the 19th Conference on Neural Information Processing Systems*, Vancouver, British Columbia,2005.

[9] C. P. Lo, A.K.W. Yeung. Concepts and Techniques in Geographic Information Systems. (Prentice-Hall of India), New Delhi, 2002.

[10] J. Xu. Formalizing the Natural language Descriptions about the Spatial Relations between Linear Geographic Objects. *Journal of Remote Sensing*, 2007, Vol. 11, No. 2, pp.152-158.

[11] M. Egenhofer and R. Herring. Categorizing Binary Topological Relationships Between Regions, Lines, and Points in Geographical Database. *Technical Report, Department of Surveying Engineering*, University of Maine, Orono, ME, 1992.

[12] X. Y. Zhang and G. N. Lv. Natural Language Spatial Relations and Their Applications in GIS. *Geo-Information Science*, 2007, Vol19, No16, pp.77-81.

[13] X. Q. Le, C. J. Yang, W. Y. Yu. Spatial Concept Extraction Based on Spatial Semantic Role in Natural Language. *Editorial Board of Geomatics and Information Science of Wuhan University*, 2005, Vo l. 30 No. 12, pp.1100-1103.

[14] C. J. Zhang, X. Y. Zhang, W. M. Jiang, etc. Rule-based Extraction of Spatial Relations in Natural Language Text. *In Proceedings of the International Conference on Computational Intelligence and Software Engineering (CiSE*). 2009, Dec.11-13, Wuhan, China.

[15] Q. J. Shen, X. Y. Zhang, W. M. Jiang. Annotation of Spatial Relations in Natural Llanguage. *In Proceedings of the International Conference on Environmental Science and Information Application Technology*. 2009, July 4-7, Wuhan, China, pp. 418-421.

[16] Y. Luo. Study on Application of Machine learning Based on Support Vector Machine. *Southwest Jiaotong University, Doctor Degree Dissertation*, 2007.