論文の内容の要旨

論文題目： Collective Semantic Annotation for Web Text: Triple Tagging and Triple Extraction

（Web テキストへの集団による意味的アノテーション：トリプルタギングとトリプル抽出）

氏　名　　楊　潔

Semantic annotations are machine-understandable metadata attached to web resources. Semantic annotations represent information contained in text documents in a structured format which are more amenable to applications in data mining, question answering, or the Semantic Web. Considerable research has been done in the reign of semantic annotation. If we check the sources of the semantics of semantic annotations, existing studies can be classified in two categories: the “ontology-centric” class which depends on the “a-prior” vocabularies (generally known as ontologies) to annotate web text; and the recent “user-centric” class which avoids pre-defined vocabularies and allows normal web users to annotate web text with less or no constraints.

This research on “collective semantic annotation” is a user-centric annotation approach. The goal of the work is to explore how we can generate semantic annotations for web text by exploiting the strengths of both normal web users and computers. Specifically, two questions are addressed. Firstly, what user-centric support can be provided to encourage normal web users annotating web text? Secondly, how to automate the annotation process?

As the result of the first question, a user-centric annotation diagram, triple tagging diagram, is proposed. I identify eight dimensions which help us to describe annotation frameworks. Literature work is investigated in terms of the eight dimensions. The features and novelties of the triple tagging diagram are addressed. The diagram consists of three parts: the concept model which defines annotation primitives, the collaboration model which addresses the information collection and navigation possibilities, and the ontology model which provides a common definition for triple annotations so that they can be exchanged, re-used, and extended on the Web. A model evaluation is carried out, which includes both qualitative and quantitative analysis. The evaluation exhibits the expressive power and advantages of the triple tagging diagram over existing work.

Regarding the second question, I propose an interactive approach which generates semantic annotations for web text automatically. In this approach, the annotation generation problem is defined as a binary relation extraction problem. Linguistics and machine learning techniques are exploited to solve the problem. Specifically, we propose the algorithm of penalty tree similarity. The algorithm is an extension of tree kernels

which are widely used in the field of Information Extraction. A triple tagging corpus is created and used in experiments. The result shows that the extended tree similarity algorithm achieves better performance.

As a result of this research, a triple tagging system, Triple-Note, is implemented. It is implemented in a web-server architecture. On the client side an extension of Firefox browser is implemented to support users' annotating actions. On the server side, automatic extraction, annotation storage, and other servicing models are implemented.