Machine Discovery Today

By: Umi Laili Yuhana (伍妮)

Machine discovery is a field of study that exploits computers to perform discovery task. It focus on search in data/observation space, problem space, hypothesis space, space of instrument, space of experiment and space of conclusion. Discovery itself refer to term that the meaning is obtain knowledge or awareness of something that unknown before. In the last decade, we have a large database come from human activity around the world. It is difficult thing to find new knowledge, new concept or new problem representations in large database manually by human. Trend research today is how to include machine / computer to help human task in discovery. To do this task, the language must be readable or understandable and analyzable by machine. Some methods are applied to this task in data preprocessing, modeling, computation (learning and searching) based on SOP for machine discovery. These methods are called machine discovery method, to find new knowledge by using computer.

Language modeling, Finite State Acceptor (FSA), Finite State Transducer (FST), Expectation Maximization (EM), Viterbi, probability, log probability, forward backward training, and semantic graph are a kind of machine discovery method that can be used. Many tasks can be solved using machine discovery application that used this methods, such as discovering a linier writing order of ancient script [1], modeling and finding abnormal instances in Multi Relational Network to extract anomaly thing [2] and find interesting facts and connections in a bibliography dataset [3].

Next section will discuss about problem example in chicken-and-egg dilemma that can be solve using EM , property in semantic graph and example of information that can be described using Multi Relational Network.

I. Chicken-and-egg dilemma that can be solve using EM

EM (Expectation Maximization) can be applied to resolve the chicken-and-egg dilemma such as language generation with incomplete data. Web document tagging is another problem that can be solved using EM. Tags tell us about the web document and vice versa. Suppose we have several tagged documents and un-tagged documents. We should give the tags to all documents. Tags are actually come from bag of words from the documents. Before web documents are tagged, they should be classified using keyword or tags for the web document. So to tag the web document we need tags. This problem is chicken-and-egg dilemma. In practice, the E-step corresponds to perform classification of each un-tagged document and calculate probability for the tags for untagged web document. M-step corresponds to calculating a new maximum a posteriori estimate for the parameters with the current estimation. E-step and M-step will be processed until the probability converged, it means that all tags have assigned to all documents with highly accurate.

II. Property of a semantic graph that might be interesting

Social network is a social structure made of nodes (which are generally individuals or organizations) that are tied by one or more specific types of interdependency. In other hand, semantic graph is often used as a form of knowledge representation. It consists of nodes that represent concept and edge, which represent semantic relations between the concepts. Semantic graph connects everything; people,

emails, products, services, web pages, documents, multimedia, groups, events, projects, activities, interests, places, companies, et cetera. Social network is a kind of semantic graph that connects the people. In other definition, semantic graph, social network and multi relational network is the same, refers to graph with labeled links and nodes.

Social network or semantic graph has several features or properties that appear to be common to networks such the small-world effect, transitivity or clustering, degree distributions, network resilience, mixing patterns, degree correlations, community structure, network motif and centrality.

Similar information in network can be used as feature to cluster people behavior in her group. As an example, people will buy product "A" if her friends in her group also buy product "A". Similarity of information can be used to clustering the people to be one group. This feature is interesting to extract all the behavior of a people or a group people, also extract the anomaly behavior of a people in her group.

III. Another type of data or information that can be represented in MRN format

Multi Relational Network (MRN), also is called as semantic graph, is a type of network that represents nodes as object of different types and binary relationships between those objects as edge/link. Many data / information (such: bibliography, movie network, film actor, email message, social contact, blogosphere) can be represented using this network. By representing information in semantic graph, it can be analyzed by machine using discovery method to discover hidden knowledge inside.

Other Information that can be described using MRN is information about photo in flickr (www.flickr.com). Photos are uploaded by user / people. Each photo has tag that given by user who upload the photo. User in www.flickr.com has email address and friend (another user). Every user can give any comment to another user's photo. And still there is much information stored in this site. Figure 1 shows Multi Relational Network that represents this information in flickr website. In this MRN, we only give 8 types of link example: tagIn, uploadBy, hasTag, friendOf, takePictureOn, commentOn, hasCommend, and hasEmail. In flickr website, there are many types link, more than 8.

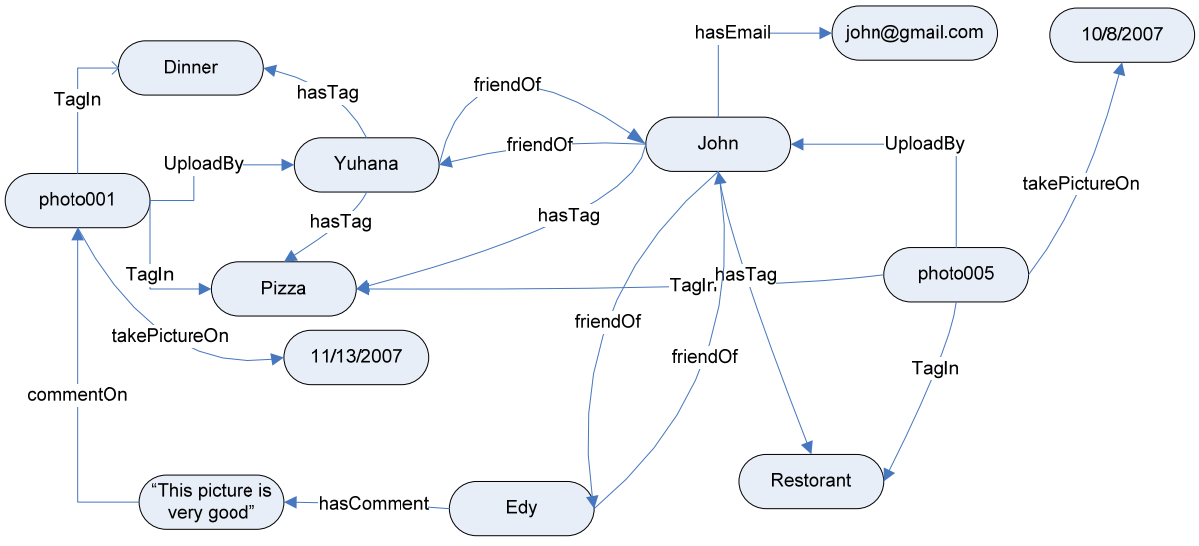

Figure 1. MRN about information in Flickr website

Using machine discovery method are described above, machine discovery can achieve its goal for helping human to find new knowledge, new problem representations, and new concepts using machine / computer. In this decade, searching and finding information are growing fast, more efficient than last decade, but still help not much. In few years later, I am sure that machine will do human task in searching and finding semantic information, anomaly things, or finding new concepts efficiently and effectively more than that human can do.

Reference:

1. Shou-de Lin and Kevin Knight "Discovering the linear writing order of a two-dimensional ancient hieroglyphic script" in Artificial Intelligence v.170/4-5, Elsevier, 2006.
2. Shou-de Lin "Modeling, Finding and Explaining Abnormal Instances in Multi-Relational Networks", Ph.D. Thesis.
3. Shou-de Lin and Hans Chalupsky "Using Unsupervised Link Discovery Methods to Find Interesting Facts and Connections in a Bibliography Dataset" *selected to be the 2nd place for the open task in* ***ACM KDDCup 2003****, in KDD Explorations V5 Issue 2.*
4. *www.flickr.Com.*
5. Machine Discovery course materials and reading list in CSIE, National Taiwan University, 2007-2008.