Information Retrieval by Linkage Discovery

Richard S. Segall
Arkansas State University, USA

Shen Lu
Soft Challenge LLC, USA

INTRODUCTION

This article discusses the topic of information retrieval by linkage discovery and reviews the related work of others in this area. Linkage discovery has been proven to be a useful method of relating sections of text, themes, and subtopics. This article discusses the relationship of linkage discovery and information retrieval including its background and a summary of work by the authors Lu and Segall (2013; 2011) and Lu et al. (2012; 2011) and that of others in the related areas of algorithms and models, multi-document summarization, web linkage and similarity measures, and linkage and semantic analysis.

With the development of richness of information using software and Internet, the need of knowledge discovery from a huge amount of information is increasing. Linkage discovery is about how to find matching records or duplicates among entities and sections within or across the files. There are many applications in linkage discovery, such as entity resolution for wrong spelling, entity resolution for data from different database systems with different data structures, entity resolution for timely changes of personal information, discovery of linkage between different sections in electric publications.

For entity resolution, the Fellegi-Sunter model can be used to discover linkage as shown in new model Entity Resolution for Fellegi-Sunter (ERFS) by Lu and Segall (2013). To discover linkage between different sections in digital publications, one can use semantic analysis. In this article, we discuss a review of literature related to the Fellegi-Sunter model, Stanford Entity Resolution Framework (SERF), Expectation Maximization (EM) and Latent Semantic Analysis (LSA) and other methods, each of which can be used for information retrieval to discover linkage from entities and sections within or across the files.

BACKGROUND

This section discusses the concepts of information retrieval, knowledge extraction, and record linkage (RL), and as well as the foundations of record matching and linkage. The latter also includes parameter estimation and knowledge discovery involving comparisons of semantic similarity between pieces of textual information within and among documents.

Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources with searches that can be based on metadata or on full-text indexing. Information retrieval can lead to new knowledge or knowledge discovery by knowledge extraction. (Wikipedia, 2012a)

Knowledge extraction is the creation of knowledge from structured (relational databases, XML) and unstructured (text, documents, images) sources. The resulting knowledge needs to be in a machine-readable and machine-interpretable format and must represent knowledge in a manner that facilitates inferencing. (Wikipedia, 2012b)

Record linkage (RL) refers to the task of finding records in a data set that refer to the same entity across different data sources Record linkage is necessary when joining data sets based on entities that may or may not share a common identifier as may be the case due to differences in record shape, storage location, and/or curator style or preference. A data set that has undergone RL-oriented reconciliation may be referred to as being *cross-linked*. (Wikipedia (2012c)]

DOI: 10.4018/978-1-4666-5888-2.ch387

Fellegi and Sunter (1969) provided a statistical model for record linkage and discussed different solutions associated with this model for different situations. They concluded that linkage rules can be defined with the observed data. With linkage rules, we can determine if a pair of records is a link, a non-link, or a possible link.

Stanford Entity Resolution Framework (SERF) (2009) provided a general framework for when and how to identify and match a pair of records. Stanford Entity Resolution Framework (SERF) is a linkage process which can be used to match and merge records, and it includes two steps: one is record matching, the other is record merging. In the matching process, it defines a black-box mechanism. All of the record pairs go through black-box and each record pair gets similarity values for different attributes. In the merging process, similar records are merged into one.

Latent Semantic Analysis (LSA) in Deerwester et al. (1990) is a general theory of acquired similarity and knowledge representation. LSA can be used to discover knowledge from text with a general mathematical learning method without knowing prior linguistic or perceptual similarity knowledge. The motivation of LSA in terms of psychology is that people learn knowledge only from similarity of individual words taken as units, not with knowledge of their syntactical or grammatical function. LSA assumes that the dimensionality of the context in which all of the local words are represented is of great importance and the reduction of dimensions of the observed data from original text to a much small but still large number can improve human cognition.

Expectation Maximization (EM) has been used in a variety of situations for parameter estimation of record linkage and is discussed in Dempster et al (1977) and Winkler (1993) as an iterative technique for estimating the value of some unknown quantity, given the values of some correlated, known quantity. It is frequently used for data clustering in machine learning. This approach is first to assume the quantity is represented as a value in some parameterized probability distribution. For parameter estimation, one can use either frequency-based parameter estimation or Expectation Maximization (EM)-based parameter estimation. The EM algorithm converges to unique limiting solutions over different starting points and is numerically stable.

Table 1 below compares information retrieval techniques of identification, connectivity measures, web page relationships, and linkage with different entities for several of the articles discussed in the following sections of this article.

INFORMATION RETRIEVAL BY LINKAGE DISCOVERY

Research by Authors: Lu and Segall

Lu and Segall (2013) used the Fellegi-Sunter model to improve the results of semantic analysis for identification of similar records. According to Lu and Segall (2013) experimental results for a new model named Entity Resolution for Fellegi-Sunter (ERFS) yielded rates of correct record classification that are higher for about 11.07% of the experiments than those using the SERF (Stanford Entity Resolution Framework.

Lu et al. (2012) discuss the operation of developing domain specific semantic space within document by inserting definitions of the terms of domain glossaries into the words of the documents. This enhances the ability to discover linkage between different sections by providing background knowledge to the text, which can improve the accuracy of context based linkage discovery.

Lu et al. (2011) discussed the use of Latent Semantic Analysis (LSA) and Expectation- Maximization (EM) to discover connections between papers within a symposium proceeding and then link papers along a variety of themes. The intention of Lu et al. (2011) was to enhance the scientific discovery process by bringing about an awareness of the relationship between contributions of different authors whose submissions and research may have had no prior relevance.

Lu and Segall (2011) discussed the linkage in medical records and bioinformatics data. In medical management systems, patients usually have multiple records for different visits. It is a general operation that different records about the same entity (person) should be merged together. Lu and Segall (2011) introduced the main techniques which are generally used to match and merge similar records, and discussed the advantage and disadvantage of those techniques. A new algorithm developed by Lu and Segall (2011), called ERFS (Entity Resolution for Fellegi-Sunter) algorithm, was provided to solve those problems for which experimental results were shown to be better