2015

Information Retrieval by Linkage Discovery

Richard S. Segall
Arkansas State University - Jonesboro, rsegall@astate.edu

Shen Lu
Soft Challenge LLC

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

Table 1. Information retrieval techniques of articles discussed in this article

Information Retrieval Technique	Algorithms and Models	Multi-document Summarization	Web Linkage and Similarity Measures	Linkage and Semantic Analysis
Identification	Identification of paraphrases [Barzilay and McKeown (2001)]	Document cluster score [Yih et al.(2007)]	Sentence similarity measures for web page retrieval [Li et al. (2006)]	Content-Assessment Module (CAM) [Bailey and Meurers (2008)]
Connectivity Measures	Connectivity analysis [Bharat and Henzinger (1998)]	Automatic syntactic simplification for content selection [Siddharthan et al. (2004)]	Shortened distances between nodes in web graph [Bjorneborn (2001)]	Latent Semantic Analysis with Expectation Maximization [Lu et al. (2011)]
Web Page Relationships	Web page relationships [Hou and Zhang (2003)]	Multilingual document clustering [Evans and Klavans (2003)]	Similarity measures for plagiarism detection [Lyon et al. (2001)]	Latent Semantic Analysis with Glossaries [Lu et al.(2012)]
Linkage with different entities	Entity Resolution for Fellegi-Sunter (ERFS) [Lu and Segall (2014)]	Different document types [McKeown et al. (2001)]	Linkage patterns among different web services [AbuJarour and Award (2011)	Unsupervised asymmetrical paraphrase detection [Joao et al. (2007)]
Other	Bayesian sentence topic model [Wang et al. (2009)]	Topic segmentation and link detection [Ferret (2002)]	Finding relevant web pages from linkage information [Hou and Zhang (2003)]	Automatic text decomposition using text themes [Salton et al. (1996)]

than those obtained using Stanford Entity Resolution Framework (SERF) that was described in more depth in Lu and Segall (2013).

Algorithms and Models

Bharat and Henzinger (1998) presented improved algorithms for topic distillation in a hyperlinked environment by studying connectivity analysis within a topic specific graph of hyperlinked documents. Bharat and Henzinger (1998) approach was to augment a previous connectivity analysis algorithm with content analysis so that a typical user query would find quality documents related to the query topic. The results of Bharat and Henzinger (1998) showed an improvement of precision for 10 documents by at least 45% over pure connectivity analysis.

Hou and Zhang (2003) presented two hyperlink analysis-based algorithms to find relevant pages for a given web page. The first algorithm of Hou and Zhang (2003) comes from the extended cocitation analysis of the web pages that is intuitive and easy to use. The second algorithm of Hou and Zhang (2003) uses linear algebra theories to reveal deeper relationships among the web pages and to identify relevant pages more precisely and effectively. Hou and Zhang (2003) showed that these algorithms could be used for various web applications, such as enhancing web searches.

Barzilay and McKeown (2001) presented an unsupervised learning algorithm for identification of paraphrases from a corpus of multiple English translations of the same source text. The approach of Barzilay and McKeown (2001) yielded phrasal and single-word lexical paraphrases as well as syntactic paraphrases. The algorithm of Barzilay and McKeown (2001) produced 9483 pairs of lexical paraphrases and 25 morpho-syntactic rules and demonstrated improved precision over methods used by other investigators, and in experiments with random 500 paraphrasing pairs of two independent judges yielded agreements of 87.8% and 91.4%.

Multi-Document Summarization

Ferret (2002) presented a method, called TOPICOLL, for using collocations for topic segmentation and link detection.

Yih et al. (2007) showed and demonstrated with successful results that a simple procedure based on maximizing the number of informative content-words

can produce some of the best reported results for multi-document summarization. Yih et al. (2007) first assigned a score to each term in the document cluster, using only frequency and position information, and then found the set of sentences in the document cluster that maximizes the sum of these scores, subject to length constraints.

Siddharthan et al. (2004) explored the use of automatic syntactic simplification for improving content selection in multi-document summarization. Siddharthan et al. (2004) showed how simplifying parenthetical by removing clauses and appositives results in improved sentence clustering and consequently better summarization, by forcing clustering based on central rather than background information.

Evans and Klavans (2003) developed a multilingual version of Columbia Newsblaster as a testbed for multilingual multi-document summarization that collects, clusters and summarizes news documents from sources around the world daily. The test bed developed by Evans and Klavans (2003) provides a platform for testing different strategies for multilingual document clustering, and approaches for multilingual multi-document summarization.

McKeown et al. (2001) discussed an in-depth approach and evaluation of Columbia multi-document summarization that covers the different component summarizers that handle different document types, the router that decides which summarizer to use, and a preliminary analysis of evaluations results relative to other systems and of factors such as the document types and the model summaries that affect the evaluation.

Wang et al. (2009) proposed a new Bayesian sentence-based topic model for summarization by making use of both the term-document and term-sentence associations, and provided experimental results on benchmark data sets that show the effectiveness of the proposed model for the multi-document summarization task.

Web Linkage and Similarity Measures

AbuJarour and Awad (2011) discovered linkage patterns among web services using business process knowledge. The main contributions of AbuJarour and Awad (2011) are finding realistic linkage patterns among web services with fine-grained types and weights, providing sources to rank recommended web services, and disambiguating exclusive relations between web services using lexical ontologies (e.g. WordNet).

Li et al. (2006) investigated sentence similarity based on semantic nets and corpus statistics by noting that sentence similarity measures play an increasingly important role in text-related research and applications in areas such as text mining, web page retrieval and dialogue systems. Li et al. (2006) focused on computing the similarity between very short texts of sentence length, and presented an algorithm that takes account of semantic information and word order information implied in sentences. According to Li et al. (2006), the semantic similarity of two sentences is calculated using information from a structured lexical database and from corpus statistics. Experiments in Li et al. (2006) were performed on two sets of selected sentence pairs that demonstrate that the proposed method provides a similarity measure that shows a significant correlation to human intuition.

Bjorneborn (2001) presented a project that included case studies of so-called co-linkage chains consisting of co-linking and co-linked web nodes that are analogous to bibliographic couplings and co-citations, into a context of researchers' homepages and published bookmark lists. Bjorneborn (2001) demonstrated that use of transversal links make the web more strongly connected by transforming into shorter distances between nodes in the web graph.

Lyon et al. (2001) investigated detecting short passages of similar text in large document collections by presenting a principle of "fingerprint extraction" in which a large collection of independently written documents each text is associated with a fingerprint which should be different form all the others. The principle underlying the system of Lyon et al. (2001) is that the identifying fingerprint associated with a piece of text is based on a large number of small, easily extracted lexical features called "word trigrams." The system of Lyon et al. (2001) was successfully used to detect plagiarism in students' work by finding small sections that are similar as well as those that are identical.

Linkage and Semantic Analysis

Bailey and Meurers (2008) diagnosed meaning errors in short answers to reading comprehension questions by developing a Content-Assessment Module (CAM) which performs shallow semantic analysis to diagnose meaning errors. The experiments conducted by Bailey

and Meurers (2008) reached an accuracy of 88% for semantic error detection and 87% on semantic error diagnosis on a data set that consisted of 566 responses to short-answer comprehension questions.

Joao et al. (2007) proposed a new type of mathematical functions for unsupervised detection of paraphrases, and tested it over a set of standard paraphrase corpora with results that outperformed state-of-art functions developed for similar tasks. Joao et al. (2007) test their new functions for both symmetrical and asymmetrical paraphrases, where the later is a pair of sentences where at least one sentence is more general or contains more information than the other.

Salton et al. (1996) introduced two main text decomposition strategies that are a chronological decomposition into text segments, and a semantic decomposition into text themes, Salton et al. (1996) then used the interaction between text segments and text themes to characterize text structure and to formulate specifications for information retrieval, text transversal and text summarization.

FUTURE RESEARCH DIRECTIONS

The future direction of this research is for the authors to perform additional experiments that demonstrate the efficiency of new methods for retrieval of information using record and web linkage. This would entail the development of new algorithms, models, and measures of information retrieval and efficiency by linkage discovery.

The future directions of the research is also to continuously update and expand the knowledge domain of the work of other investigators in the area of information retrieval by linkage discovery, and to provide an extended synthesis of categorization of the information retrieval techniques such as expanding that as shown in Table 1.

CONCLUSION

This article illustrates that the topic of information retrieval by linkage discovery is an active area of research investigation that is useful for both the large scale domain of the web and large databases of documents and records.

Knowledge discovery and extraction from structured and unstructured text sources, summarization from multi-documents, web linkage and similarity measures, and linkage and semantic analysis are just some of the areas of linkage discovery that are useful for the ever expanding area of information retrieval.

REFERENCES

AbuJarour, M., & Awad, A. (2011). Discovering linkage patterns among web services using business process knowledge. In *Proceedings of the 2011 IEEE International Conference on Services Computing* (pp. 314-321). Washington, DC.

Bailey, S., & Meurers, D. (2008). Diagnosing meaning errors in short answers to reading comprehensive questions. In *Proceedings of the Third ACL Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 107-115).

Barzilay, R., & McKoewn, K. R. (2001). Extracting paraphrases from a parallel corpus. In ACL '01 *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics* (pp. 50-57).

Bharat, K., & Henzinger, M. R. (1998). Improved algorithms for topic distillation in a hyperlinked environment. In *SIGIR'98, Proceedings of the 21st Annual International ACM Special Interest Group on Information Retrieval (SIGIR) Conference on Research and Development in Information Retrieva*l (pp. 104-111). Melbourne, Australia.

Bjorneborn, L. (2001). Small-world Linkage and co-linkage. In *Proceedings of the 12th ACM conference on Hypertext and Hypermedia* (HYPERTEXT '01) (pp. 133-137).

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science American Society for Information Science*, *41*(6), 391–407. doi:10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B. Methodological*, *39*(1), 1–38. Retrieved from http://babbage.cs.missouri.edu/~chengji/mlbioinfo/CiteULike

Evans, D. K., & Klavans, J. L. (2003). A platform for multilingual news summarization. *Columbia University, Department of Computer Science*. Retrieved April 20 from http://www.cs.columbia.edu/~library/TR-repository/reports/reports-2003/cucs-014-03.pdf

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*, 1183–1210. doi:10.1080/01621459.1969.10501049

Ferret, O. (2002). Using collocations for topic segmentation and link detection. In *Proceedings of the 19th International Conference on Computational Linguistics (LOLING)* (v. 1 pp. 1-7). Association of Computational Linguistics.

Hatzivassiloglou, V., Klavans, J. L., Holcombe, M. L., Barzilay, R., Kan, M.-Y., & McKoewn, K. R. (2001). SimFinder: A flexible clustering tool for summarization. In *Proceedings of the NAACL Workshop on Automatic Summarization* (v. 1). Association for Computational Linguistics.

Hou, J., & Zhang, Y. (2003). Effectively finding relevant web pages from linkage information. *IEEE Transactions on Knowledge and Data Engineering*, *15*(4), 940–951. doi:10.1109/TKDE.2003.1209010

Joao, C., Gael, D., & Pavel, B. (2007). New functions for unsupervised asymmetrical paraphrase detection. *Journal of Software*, *2*(4), 12–23. doi:10.4304/jsw.2.4.12-23

Li, Y., McLean, D., Bandar, Z., O'Shea, J. D., & Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering*, *18*(8), 1138–1150. doi:10.1109/TKDE.2006.130

Lu, S., & Segall, R. S. (2011). Linkage in medical records and bioinformatics data. In *Proceedings of the 2011 MidSouth Computational Biology and Bioinformatics Society (MCBIOS) Meeting*. College Station, TX.

Lu, S., & Segall, R. S. (2013). Linkage in medical records and bioinformatics data. [IJIDS]. *International Journal of Information and Decision Sciences*, *5*(2), 169–187. doi:10.1504/IJIDS.2013.053803

Lu, S., Segall, R. S., & Belford, R. E. (2011). Linkage discovery with symposium proceedings. To appear in *Proceedings of the 6th BioNanoTox and Applications International Research Conference*. Little Rock, AR. Retrieved from http://www.softchallenge.net/sites/default/files/LU_SEGALL_BELFORD_bionanotox2011.pdf

Lu, S., Segall, R. S., & Belford, R. E. (2012). Linkage discovery by combining latent semantic analysis with glossaries from biosciences domains. *2012 Arkansas NSF EPSCoR Annual Meeting*. Springdale, AR.

Lyon, C., Malcolm, J., & Dickerson, B. (2001). Detecting short passages of similar text in large document collections. In *Proceedings of Conference on Empirical Methods in Natural Language Processing* (pp. 118-125). Pennsylvania.

McKoewn, K. R., Barzilay, R., & Evans, D. Hatzivassiloglou, V., Teufel, S., Kan, Y. M, Schiffman, B. (2001). Columbia multi-document summarization: Approach and evaluation. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Development.* New Orleans, LA.

Perkowitz, M., & Etzioni, O. (1998). Adaptive web sites: Automatically synthesizing web pages. *'AAAI '98/IAAI '98: Proceedings of the 15th National/10th Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence'*, American Association for Artificial Intelligence (pp. 727-732). Menlo Park, CA, USA.

Salton, G., Singhal, A., Buckley, C., & Mitra, M. (1996). Automatic text decomposition using text segments and text themes. [Washington, DC.]. *Hypertext*, *96*, 53–65.

Siddharthanm, A., Nenkova, A., & McKeown, K. (2004). Syntactic simplification for improving content selection in multi-documentation summarization. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)* (pp. 896). Retrieved from http://acl.ldc.upenn.edu/coling2004/MAIN/pdf/129-828.pdf.

Stanford University InfoLab. (2009). Stanford Entity Resolution Framework. Stanford University, Stanford, CA. Retrieved from http://infolab.stanford.edu/seaf.

Wang, D., Zhu, S., Li, T., & Gog, Y. (2009). Multi-Document Summarization using Sentence-based Topic Methods. In ACLShort '09, *Proceedings of the ACL-IJCNLP Conference Short Papers* (pp. 297-300).

Wikipedia. (2012a). Information Retrieval. Retrieved from http://en.wikipedia.org/wiki/Information_retrieval.

Wikipedia. (2012b). Knowledge Extraction. Retrieved from http://en.wikipedia.org/wiki/Knowledge_extraction.

Wikipedia. (2012c). Record Linkage. Retrieved from http://en.wikipedia.org/wiki/Record_linkage.

Wikipedia. (2012d). Semantic Analysis. Retrieved from http://en.wikipedia.org/wiki/Semantic_analysis_(linguistics).

Wikipedia. (2012e). Semantic Similarity. Retrieved from http://en.wikipedia.org/wiki/Semantic_similarity.

Wikipedia. (2013). Topic Segmentation. Retrieved from http://en.wikipedia.org/wiki/Topic_segmentation.

Winkler, W. E. (1993). Improved decision rules in the Felligi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods* (pp. 274-279). American Statistical Association. Retrieved from http://www.cenus.gov/srd/papers/pdf/rr93-12.pdf.

Winkler, W. E. (1999). The state of record linkage and current research problems. *Technical Report, Statistical Research Division*. Washington, DC, USA: US Bureau of the Census. Retrieved from http://www.cenus.gov/srd/papers/pdf/rr99-04.pdf.

Yih, W.-T., Goodman, J., Vanderwende, L., & Suzuki, H. (2007). Multi-document summarization by maximizing informative content-words. In *IJCAI'07 Proceedings of the 20th international joint conference on Artificial intelligence* (pp. 1776-1782).

ADDITIONAL READING

Baxter, R., & Christen, P. (2003). A comparison of fast blocking methods for record. *First Workshop on Data Cleaning, Record Linkage and Object Consolidation*, Knowledge Discovery in Data (KDD), Washington, DC, USA, available at http://www.datamining.anu.edu.au/./publications/203/k.

Brkowski, W., & Mielniczuk, H. (2003). Use of medical record linkage in epidemiological studies. *Przeglad Epidemiologiczny*, *57*(3), 553–559. PMID:14682175

Cohen, W. W. (2003a). Probabilistic record linage: a short tutorial. Available at http://www.cs.cmu.edu/~wcohen/Matching-1.ppt.

Cohen, W. W. (2003b). Record linkage tutorial: distance metrics for text. Available at http://www.cs.cmu.edu/~wcohen/Matching-2.ppt.

Cohen, W. W. (2006). Annotated bibliography for matching. Available at http://www.cs.cmu.edu/~wcohen/matching/.

Dey, D., Mookerjee, V.S. and Liu, D. (2011). Efficient techniques for online record linkage. *IEEE Transactions on Knowledge and Data Engineering*, March, Vol. 23, No. 3, pp.373–387.

Gill, L., Coldacre, M., Simmons, H., Bettley, G., and Griffith, M. (1993). Computerized linkage of medical records: methodological guidelines. *J Epidemiol Community Health*, August, Vol. 47, No. 4, pp.316–319.

Gill, L. E. (1997). *OX-LINK: the Oxford medical record linkage, record linkage techniques* (pp. 15–33). Textbook in Medical Record Linkage.

Sauleau, E. A., Paumier, J.-P., & Buemi, A. (2005). Medical record linkage in health information systems by approximate string matching and clustering. *BMC Medical Informatics and Decision Making*, *5*(32), available at http://www.biomedcentral.com/1472-6947/5/32 PMID:16219102

Segall, R. S., & Lu, S. (2014). *Linkage Discovery with Glossaries, to appear in Encyclopedia of Business Analytics & Optimization, John Wang* (I. G. I. Global, Ed.). Hersey, PA.

Winkler, W. E. (1994). *Advanced methods for record linkage. Statistical Research Report Series No. RR1994/05, Statistical Research Division, Methodology and Standards Directorate*. Washington, DC, USA: US Bureau of Census.

Winkler, W. E. (2000). *Frequency-based matching in Fellegi-Sunter model of record linkage. Statistical Research Report Series No. RR2000/06, Statistical Research Division, Methodology and Standards Directorate*. Washington, DC, USA: US Bureau of the Census.

Winkler, W. E. (2006). *Overview of record linkage and current research directions. Statistical Research Report Series No. RR2006/03, Statistical Research Division, Methodology and Standards Directorate*. Washington, DC, USA: US Census Bureau.

KEY TERMS AND DEFINITIONS

Information Retrieval: The activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text indexing.

Latent Semantic Analysis: Statistical model of word usage that permits comparisons of semantic similarity between pieces of textual information (Foltz, 1996).

Linkage Discovery: Discovery of connections between or among related topic segments located in different sections of electronic publications.

Record Linkage: How to find matching records or duplicates among entities and sections within or across files.

Semantic Analysis: the process of relating syntactic structures, from the levels of phrases, clauses, sentences and paragraphs to the level of the writing as a whole, to their language-independent meanings (Wikipedia, 2012d).

Semantic Similarity: A concept whereby a set of documents or terms within term lists are assigned a metric based on the likeness of their meaning/semantic content (Wikipedia, 2012e).

Topic or Text Segmentation: The process of dividing written text into meaningful units, such as words, sentences or topics (Wikipedia, 2013).