Overview of reference paper for Project 4

Jing Wu and Tian Zheng

1 Problem formulation

1. **Author name disambiguation** is the problem of determining whether records in a publications database refer to the same person.

2. Other related topics:

   - **Entity resolution**: Entity resolution refers to problem of identifying and linking/-grouping different manifestations of the same real world object. Entity resolution operates on natural language text; author name disambiguation operates primarily on metadata about authors and articles. Examples of manifestations and objects:
     - Different ways of addressing (names, email addresses, Facebook accounts) the same person in text.
     - Web pages with differing descriptions of the same business.
     - Different photos of the same object.
   - **Record linkage**: Record linkage refers to the process by which one identifies multiple records in a database, or across two different databases, as referring to the same individual.

3. Two major challenges in author name disambiguation:

   - **synonyms**: A person may have multiple names. For example, a person named David S. Johnson can write his name as David S. Johnson, David Johnson, D. S. Johnson, or D. Johnson, etc.
   - **homonyms**: Different persons may share the same name. For example, there may exist two or more different persons named David S. Johnson.

   In this project, we focus on the second challenge - homonyms. By looking at authors appearing in the domain of bibliography, and primarily utilizing domain-specific features such as co-authorship, titles of articles, titles of publication, topics of articles, etc, you are expected to address the resolution of homonymous author names appearing in citation data.

4. Scientific tasks behind author name disambiguation (Smalheiser and Torvik [2009])

   - Investigators who are searching for potential collaborators in different disciplines seek authors (and not just their papers) because individuals are a great source of unpublished information - ideas, raw data, failed experiments, and hypotheses that were never followed up.
   - Author name disambiguation can also be generalized to any kind of person searching. For example, finding homepages of query names. When a person name is submitted to a web, people search system that finds persons in web pages, the retrieved pages may contain numerous homonymous names of the input query, decreasing the precision of the search system.

- Knowing individuals (not merely author names) is crucial for establishing new resources such as citation networks, collaboration networks, and author profiles.

2 Dataset description

1. The dataset is downloaded from http://clgiles.ist.psu.edu/data/
2. There are 14 *.txt* files in the data folder. Each file is a collection of ambiguous names and associtated citations. e.g. AGupta.txt is the citation files of 26 "A. Gupta"s. The 14 canonical names are top ranked ambiguous names, such as "J. Lee", "J Smith", "S. Lee"and "Y. Chen" from the DBLP bibliography.
3. For evaluation, the author of the dataset carefully manually labeled the canonical name entities and associated citations.
4. The datasets are pre-processed as follows. All the author names in the citations were simplified to first name initial and last name. For example, "Yong-Jik Kim" was simplified to "Y. Kim". A reason for such simplification is that the first name initial and last name format is popular in citation records. Publication dates are eliminate from citations.
5. All citations in the raw data are in the format of

   clusterid_citationid authors;authors;...<>paper title<>publication venue title,

   where *clusterid* indicates the canonical author id.
6. The ideal form of dataframe for furthur analysis is shown in Table 1.

Table 1: Author name disambiguation example

AuthorID	PaperNo	QuestAuthor	Coauthor	Paper	Journal
12	1	A Kumar	Vishv M Malhotra	A Look-Ahead Interpreter for Sequential Prolog and Its Implementation	FSTTCS Foundations of Software Technology and Theoretical Computer Science
12	2	A Kumar		A New Computation Rule for Prolog	Inf Process Lett
13	1	A Kumar	Manish Kumar Shukla ; P C Mishra ; S K Mishra	An ab initio study of excited states of guanine in the gas phase and aqueous media: Electronic transitions and mechanism of spectral oscillations	Journal of Computational Chemistry

3 Collection of methods

3.1 Research approaches to author name disambiguation

1. Most disambiguation approaches fall into one of the two machine learning paradigms: supervised or unsupervised.

- Supervised approaches take as input a set of training examples consisting of pairs of articles that are labeled as either positive (author match) or negative (not author match).
- Unsupervised approaches do not use labeled training examples.

2. The features that are available for prediction vary across datasets (In the dataset we provided, co-authorship, titles of articles, titles of publication are available.)
   - Most, but not all, approaches to disambiguation involve collapsing all of the feature scores into a single numeric value that indicates the degree of similarity between a pair of papers.
   - Some models transform features into sets of latent (or hidden) variables.
3. Using co-author network:
   - For instance, when one citation contains D. S. Johnson and C. J. Date as its authors, we can say that a D. S. Johnson in the citation indicates the D. S. Johnson whom C. J. Date knows. When another citation has also D. S. Johnson and C. J. Date, we can further state that "D. S. Johnson"s in the two citations are the same individual, namely, the D. S. Johnson whom C. J. Date knows, under the assumption that C. J. Date's in the two citations are not different persons.
   - One way to detect and correct these so-called transitivity violations is to look at sets of three papers at once and assess the transitivity.

3.2 Summary of potential papers

1. **On co-authorship for author disambiguation** (Kang et al. [2009])
   - Coauthor disambiguation hypothesis: the identity of an author is characterized by his/her coauthors.
   - Using single-link agglomerative clustering to group papers having the same name author appearances.
2. **Two supervised learning approaches for name disambiguation in author citations** (Han et al. [2004])
   - Two supervised methods are proposed based on Naive Bayes and Support Vector Machines. The methods learn a specific model for each author name from the train data and use the model to predict whom a new citation is authored by.
3. **Name disambiguation in author citations using a k-way spectral clustering method** (Giles et al. [2005])
   - Propose an unsupervised learning approach using K-way spectral clustering method. They calculate a Gram matrix for each person name and apply K way spectral clustering algorithm to the Gram matrix to get the result.
   - Spectral clustering methods compute eigenvalues and eigenvectors of a Laplacian matrix (or singular values and singular vectors of certain matrix) related to the given graph, and construct data clusters based on such spectral information.

- Another reference to better understand section 3.3: https://papers.nips.cc/paper/1992-spectral-relaxation-for-k-means-clustering.pdf

4. **Efficient topic-based unsupervised name disambiguation** (Song et al. [2007])
   - Two-stage approach
   - During the first stage, the paper presents two topic-based models inspired by two generative models for documents: Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA).
   - At the second stage, person names are disambiguated by leveraging an unsupervised hierarchical agglomerative clustering method.
5. **Author disambiguation using error-driven machine learning with a ranking loss function** (Culotta et al. [2007])
   - The paper proposes a training algorithm that is (1) error-driven in that training examples are generated from incorrect predictions on the training data, and (2) rankbased in that the classifier induces a ranking over candidate predictions.
   - There are different detailed methods in each component of the algorithm: clusterwise scoring functions, error-driven example generation, and rank-based training, shown in Table 3 in the paper.
6. **A constraint-based probabilistic framework for name disambiguation** (Zhang et al. [2007])
   - This paper first gives a constraint-based probabilistic model for semi-supervised name disambiguation.
   - Employing EM algorithm to learn the parameters of the distance measure.
   - The paper defines six types of constraints. You should clarify which constraint can be used according to the data provided.

References

Neil R Smalheiser and Vetle I Torvik. Author name disambiguation. *Annual review of information science and technology*, 43(1):1–43, 2009.

In-Su Kang, Seung-Hoon Na, Seungwoo Lee, Hanmin Jung, Pyung Kim, Won-Kyung Sung, and Jong-Hyeok Lee. On co-authorship for author disambiguation. *Information Processing & Management*, 45(1):84–97, 2009.

Hui Han, Lee Giles, Hongyuan Zha, Cheng Li, and Kostas Tsioutsiouliklis. Two supervised learning approaches for name disambiguation in author citations. In *Digital Libraries, 2004. Proceedings of the 2004 joint ACM/IEEE conference on*, pages 296–305. IEEE, 2004.

C Lee Giles, Hongyuan Zha, and Hui Han. Name disambiguation in author citations using a k-way spectral clustering method. In *Digital Libraries, 2005. JCDL'05. Proceedings of the 5th ACM/IEEE-CS Joint Conference on*, pages 334–343. IEEE, 2005.

Yang Song, Jian Huang, Isaac G Councill, Jia Li, and C Lee Giles. Efficient topic-based unsupervised name disambiguation. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 342–351. ACM, 2007.

Aron Culotta, Pallika Kanani, Robert Hall, Michael Wick, and Andrew McCallum. Author disambiguation using error-driven machine learning with a ranking loss function. 2007.

Duo Zhang, Jie Tang, Juanzi Li, and Kehong Wang. A constraint-based probabilistic framework for name disambiguation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 1019–1022. ACM, 2007.