New system solves the 'who is J. Smith' puzzle

December 14 2006

Penn State researchers have developed an automated system that can determine which "J. Smith" is authoring papers on computer science—the one who teaches at Penn State or the one who teaches at M.I.T—as well as whether "J. Smith" is John Smith, Jane Smith, Joanna L. Smith or James H. Smith.

The system, which retrieves classes of authors with similar names, considers not just names in making its determination but also other information such as co-authors, dates of publications, citations and keywords.

When tested with 3,355 academic papers written by 490 authors, the system correctly identified authors 90.6 percent of the time.

"It works very similarly to how humans would figure out authors' identity—by looking at affiliations, topics, publications," said C. Lee Giles, the David Reese Professor of Information Sciences and Technology and principal researcher.

"The system works by using machine-learning methods to cluster together names that the system believes to be similar. If you think there's another parameter that's relevant, you can change the algorithm and include it," Giles said.

The system is explained in a paper, "Efficient Name Disambiguation for Large-Scale Databases," presented at the recent 17th European

Conference on Machine Learning and the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases in Berlin. Co-authors were Jian Huang, a doctoral student in the College of Information Sciences and Technology, and Seyda Ertekin, a doctoral student in the Department of Computer Science and Engineering. Even in academic publications, figuring out an author's identity can be difficult as publications vary in how individuals' names are presented. For instance, some publications opt just for first initial and last name as in "J. Smith." Others include full name—C. Lee Giles, for instance. But if the surname is common, as in "Smith" or "Chen," first names may not suffice to accurately identify the author.

Confusion also can occur because of how entities are listed with some publications choosing Penn State, The Pennsylvania State University or PSU. The researchers' algorithm can clear up ambiguities surrounding entities whether institutions, businesses, funding agencies or organizations.

"This method will work on many entity disambiguation problems," Giles said.

The algorithm uses a clustering method to train computers to extract information based on similar properties. Each time information is clustered, the result is a smaller and smaller grouping.

The algorithm will be a part of the next generation CiteSeer, the largest academic search engine for computer and information-science literature. Giles was co-creator of CiteSeer when he was at NEC.

Source: Penn State