PHYS.ORG

Training computers to differentiate between people with the same name

January 12 2017

Mohammad al Hasan, Ph.D., is an associate professor of computer science at Indiana University-Purdue University Indianapolis. Credit: Whitney Walker, School of Science, Indiana University-Purdue University Indianapolis

All individuals are unique but millions of people share names. How to distinguish—or as it is technically known, disambiguate—people with common names and determine which John Smith or Maria Garcia or Wei Zhang or Omar Ali is a specific John Smith, Maria Garcia, Wei Zhang or Omar Ali—or even someone previously unidentified?

This conundrum occurs in a wide range of environments from the bibliographic—which Anna Hernandez authored a specific study?—to the law enforcement—which Robert Jones is attempting to board an airplane flight?

Two computer scientists from the School of Science at Indiana University-Purdue University Indianapolis and a Purdue University doctoral student have developed a novel-machine learning method to provide better solutions to this perplexing problem. They report that the new method is an improvement on currently existing approaches of name disambiguation because the IUPUI method works on streaming data that enables the identification of previously unencountered John Smiths, Maria Garcias, Wei Zhangs and Omar Alis.

Existing methods can disambiguate an individual only if the person's records are present in machine-learning training data, whereas the new method can perform non-exhaustive classification so that it can detect the fact that a new record which appears in streaming data actually belongs to a fourth John Smith, even if the training data has records of only three different John Smiths. "Non-exhaustiveness" is a very important aspect for name disambiguation because training data can never be exhaustive, because it is impossible to include records of all living John Smiths.

"Bayesian Non-Exhaustive Classification—A Case Study: Online Name Disambiguation usingTemporal Record Streams" by Baichuan Zhang, Murat Dundar and Mohammad al Hasan is published in Proceedings of the 25th International Conference on Information and Knowledge

Management. Zhang is a Purdue graduate student. Dundar and Hasan are IUPUI associate professors of computer science and internationally respected experts in machine learning.

"We looked at a problem applicable to scientific bibliographies using features like keywords, and co-authors, but our disambiguation work has many other real-life applications—in the security field, for example," said Hasan, who led the study. "We can teach the computer to recognize names and disambiguate information accumulated from a variety of sources—Facebook, Twitter and blog posts, public records and other documents—by collecting features such as Facebook friends and keywords from people's posts using the identical algorithm. Our proposed method is scalable and will be able to group records belonging to a unique person even if thousands of people have the same name, an extremely complicated task.

"Our innovative machine-learning model can perform name disambiguation in an online setting instantaneously and, importantly, in a non-exhaustive fashion," Hasan said. " Our method grows and changes when new persons appear, enabling us to recognize the ever-growing number of individuals whose records were not previously encountered. Also, some names are more common than others, so the number of individuals sharing that name grows faster than other names. While working in non-exhaustive setting, our model automatically detects such names and adjusts the model parameters accordingly."

Machine learning employs algorithms - sets of steps—to train computers to classify records belonging to different classes. Algorithms are developed to review data, to learn patterns or features from the data, and to enable the computer to learn a model that encodes the relationship between patterns and classes so that future records can be correctly classified. In the new study, for a given name value, computers were "trained" by using records of different individuals with that name to build a model that distinguishes between individuals with that name, even individuals about whom information had not been included in the training data previously provided to the computer.

"Features" are bits of information with some degree of predictive power to define a specific individual. The researchers focused on three types of features:

1. Relational or association features to reveal persons with whom an individual is associated: for example, relatives, friends, and colleagues
2. Text features, such as keywords in documents: for example, repeated use of sports- culinary-, or terrorism-associated keywords
3. Venue features: for example, institutions, memberships or events with which an individual is currently or was formerly associated

The study was funded by the National Science Foundation through CAREER awards to Hasan and Dundar in 2012 and 2013, respectively.

The researchers hope to continue this line of inquiry, scaling up with the support of enhanced technologies, including distributed computing platforms.

Provided by Indiana University-Purdue University Indianapolis School of Science

Citation: Training computers to differentiate between people with the same name (2017, January 12) retrieved 3 May 2024 from https://phys.org/news/2017-01-differentiate-people.html