Entity Disambiguation Using Semantic Networks

Jorge H. Román
Los Alamos National Laboratory, P.O. Box 1163 (MS B295), Los Alamos, NM 87545. E-mail: jhr@lanl.gov

Kevin J. Hulin
Department of Computer Science, University of Texas at Dallas, 800 West Campbell Road, Richardson, TX 75080. E-mail: kjhulin@utdallas.edu

Linn M. Collins and James E. Powell
Los Alamos National Laboratory, P.O. Box 1163 (MS P362), Los Alamos, NM 87545. E-mail: {linn; jepowell@lanl.gov}

A major stumbling block preventing machines from understanding text is the problem of entity disambiguation. While humans find it easy to determine that a person named in one story is the same person referenced in a second story, machines rely heavily on crude heuristics such as string matching and stemming to make guesses as to whether nouns are coreferent. A key advantage that humans have over machines is the ability to mentally make connections between ideas and, based on these connections, reason how likely two entities are to be the same. Mirroring this natural thought process, we have created a prototype framework for disambiguating entities that is based on connectedness. In this article, we demonstrate it in the practical application of disambiguating authors across a large set of bibliographic records. By representing knowledge from the records as edges in a graph between a subject and an object, we believe that the problem of disambiguating entities reduces to the problem of discovering the most strongly connected nodes in a graph. The knowledge from the records comes in many different forms, such as names of people, date of publication, and themes extracted from the text of the abstract. These different types of knowledge are fused to create the graph required for disambiguation. Furthermore, the resulting graph and framework can be used for more complex operations.

In collaboration with David Izraelevitz, Miriam Blake, and Gary Grider.
Received October 13, 2011; revised February 7, 2012; accepted February 8, 2012

 DOI: 10.1002/asi.22672

Background

Fusion of different types of information is not a new challenge. A common approach is one that allows mapping of information in different forms into one representation. One such standard representation is Resource Description Framework (RDF) triple standards, which can be used to build semantic networks. RDF is a standard established by the Semantic Web (http://www.w3.org/RDF). Semantic networks also have been an area of research for many years, but recent implementations capable of supporting very large networks with billions of edges (facts in the knowledge base) make them usable for real-world problems. However, one drawback that plagues this and any other approach is that ambiguous entity references are contained in the input records, thereby needing disambiguation. Disambiguating entities in networks also leads to a more efficient and compact representation by collapsing coreferent nodes, and the resulting graph is well suited for more complex operations.

Ontologies, another area of research, have been used to generate compact networks. By using ontologies, the mapping of different information types can be made into one semantic representation. Thereby, disambiguation is explicitly done. Semantic network tools can then be used to fuse information, to reason about the knowledge in the graph, and to deduce new information. However, most semantic

network and ontology-creation processes have been driven by human experts as they try to map specific fields of interest. This process, which can be highly accurate, is time-consuming and requires detailed knowledge of the domain. It may work well on limited and small domains, but real-world problems can be very large, and they depend on input from many sources. Furthermore, the content may be constantly evolving as new sources are found or the focus of the domain shifts, which would require constant revisions and/or exceptions.

Author disambiguation also has been an area of research for many years. A literature search for concepts related to author/entity disambiguation found references back to 1978. In the early days, this problem was identified as one of the most difficult to solve in artificial intelligence. Pattern analysis and machine learning were discussed as common approaches. One of the first references to semantic networks in this area was made in an article published in 1984 as part of learning-by-example technologies (Durham, 1984); however, the emphasis of that article was on machine understanding of human language. It was part of the "Fifth Generation Fever." In the late 1980s, neural networks were shown to address this problem with limited success. Similarly, various probabilistic algorithms also have been evaluated over the years, also with limited success. In the last five years, new approaches have been evaluated in this problem domain; one of them involves building similarity functions for web page content to solve entity identification of the persons referenced on the web pages (Yerva, Miklos, & Aberer, 2010). Another approach has been to generate document clusters for a given set of names (Iosif, 2010). Semantic association is discussed as another approach where named entities are clustered based on their semantic association (Blanchon & Boitet, 2006).

Social graphs and people clusters based on their social association also have been used (Rowe, 2009). In 2006, McRae-Spencer and Shadbolt documented an approach with high accuracy rates using citation networks, coauthorship, and source analysis. In 2001, Bell and Sethi wrote a review of the different approaches as they relate to patient-record matching. In Table 1 of that article, each technique is compared to the problem area it addresses. It noted that word distance techniques are effective for letter mismatches, and rule base or fuzzy logic is well suited for metadata error matching. However, note that most of the approaches used in these studies only leverage one or two dimensions of the data. On selected sets, these approaches could yield high-accuracy results. They would be less accurate in dealing with foreign names, those names that may have alternate spellings, and more general data sets that have little or no structured data.

The Los Alamos National Laboratory (LANL) Digital Knowledge Discovery research team has focused on building and using automated tools to extract as many dimensions as possible from unstructured text (Román & Spearing, 2009b, 2009c). These tools for knowledge and feature extraction identify the dimensions of one document or record. The dimensions range from conceptual to temporal, organizational, and geographical. The tools can be used out of the box; neither training nor customization is required. The automatically derived dimensions of text can be fused (Román & Spearing, 2009a) with structured information from a record to build a graph of the knowledge contained in a set of bibliographic entries.

In this project, we take advantage of the knowledge acquired by all these research fields and attempt to take the best from each field and incorporate the algorithms into a coherent framework to prove the hypothesis.

TABLE 1. First 25 unique persons from 50-record set.

Last_name(s), First_name(s)	(Initials)	No. of Documents
aadland, ____	(a r k)	1
Abd el nabi, sami	(a e n s h)	1
Abdel raouf, ____	(a r m w)	1
abdelouas, ____	(a a)	1
abdou, ____	(a h)	1
abe, hitoshi	(a h)	2
abrefah, ____	(a j)	1
abu-eid, ____	(a e r)	1
acarkan, ____	(a s)	1
achuthan, ____	(a p v)	1
ackerman, ____	(a j p)	1
ackland, ____	(a m c)	1
adam, ____	(a e)	1
adamov, ____	(a e o)	1
aden, ____	(a v g)	1
adiwardoyo, ____	(a)	1
afanas_ev, ____	(a e a a)	1
afanasieva, ____	(a e)	1
afnasyev, ____	(a a)	1
aggeryd, ____	(a i)	1
aghara, sukesh	(a s k)	1
agostini, ____	(a p)	1
Ahmady ibrahim, ____	(a i m e)	1
ahn, ____	(a s j)	1
ahn, joon	(a j h)	5

Approach

A semantic network is the underlying information representation chosen for the approach. The framework uses several algorithms to generate subgraphs in various dimensions. For example: a person's name is mapped into a phonetic dimension, the abstract is mapped into a conceptual dimension, and the rest are mapped into other dimensions. To map a name into its phonetic representation, an algorithm translates the name of a person into a sequence of phonemes. Therefore, two names that are written differently but pronounced the same are considered to be the same in this dimension. The "same" qualification in one of these dimensions is then used to identify potential coreferent entities. Similarly, an algorithm for generating potential alternate spellings of a name has been used to find entities for comparison with similarly spelled names by computing word distance.

DOI: 10.1002/asi

A prototype framework has been created. It takes bibliographic records with unstructured text abstracts and automatically generates a semantic network where the dimensions contained in the data can be fused. From the structured data in the bibliographic records, the author(s) of a paper can be identified along with, for example, publication date, affiliation, and conference information.

The hypothesis underlying our approach is that coreferent entities are strongly connected on a well-constructed graph. However, automatically generated networks are inefficient because they contain many ambiguous entities. For example, bibliographic records often reference the same person using different labels such as full names or initials only for the first name. Therefore, entity disambiguation is crucial to generating usable networks. This area has been researched by many different groups, yet, to date, no one solution has addressed all the problems. The approach presented in this article uses a generalized framework that takes advantage of other algorithms to improve overall accuracy when identifying coreferent entities. The remainder of this article will document the specific implementation for bibliographic record entity disambiguation, but the approach is generic enough to support entity disambiguation of many different kinds, not just authors.

Data and Algorithms

The raw bibliographic records used were in Digital Item Declaration Language (DIDL) format, and the facts in each were translated into a set of RDF triples (Subject, Predicate, Object). A sample bibliographic record is composed of structured fields such as the names of the authors, the conference where it was presented, date of publication, and an abstract. The abstract contains unstructured text which summarizes the content of the publication.

Each record was processed separately and transformed into RDF triples, which map the record into the different dimensions. The key dimensions used in this case were literal, phonetic, and word distance match of author names, and the conceptual dimension for the abstract. Selected structured data fields such as affiliation, e-mail, and publication language also were imported as triples. This process generated a large network. The initial expansion generated an average of 80 triples per record.

Algorithms such as the improved Levenshtein (http://www.merriampark.com/ld.htm) distance and a phoneme decomposition (taken from http://freetts.sourceforge.net) were used to augment the graph so that the machine would be able to identify like-sounding names. These two algorithms were chosen from a variety of options tested because they seem to generate accurate mappings into the dimensions of interest. For the abstract portion of the record, a theme extraction algorithm was used. The algorithm is an automated implementation of the "speed reading" technique described by Turney in 1999 and is commercially available from CiriLab Canada (http://www.cirilab.com). It uses word frequency and other natural language traits to select important themes and the relationship to other themes as found in the text (Román et al., 2008). Themes can be multiword concepts. The algorithm requires no user input; it simply selects themes from the text. The autonomous nature of the algorithm allows the use of any data set with no predefined categories. The algorithm derives the themes and their relationships based on position and frequency as they appear in the body of the text (Collins et al., 2009).

For storage, a federated implementation of AllegroGraph (AG) Version 4.0 Server was used, and queries of these triples were quick and efficient (http://www.franz.com/). AG implements path-finding functions that are key to finding coreferences. Therefore, for path searching, the framework uses AG's Social Network Analysis functions, which can compute the shortest path and find all paths between two nodes. The path search can be restricted by length and by edge type. For example, conceptual matching takes place along "theme" edges only and looks for connectivity between two documents authored by the candidates. The connectivity identifies two publications in the same conceptual dimension.

Representation and Framework Implementation

The triples for Document 11 are shown in a graph in Figure 1. It contains automatically generated key-themes which are shown connected to related themes. It also contains references to authors which have a "pers:" prefix, conferences use "confs:", organizations use "orgs:", and general facts have no prefixes. For example, the year of publication is represented by a triple (doc11, yearPublished, 2000-05-01). Since all the facts for doc11 are shown, the node "doc11" is the center of this graph. Note that the graphical interface to the semantic network is highly interactive; it allows the user to zoom and pan as well as hide details. It also uses color, as shown in Appendix A, to denote the different dimensions of the data. Images shown are screen captures of this interface. These illustrations are meant to be indicative of the complexity of the graphs, and some labels may not be readable. Use of the interface is recommended for detailed exploration.

Each person named in an abstract is represented as a unique entity. For example, one of the authors is represented as (docs:doc11, authors, pers:pers31). The name of this person is then broken down into details: first_name, last_name, and initials. Some additional information is carried from the record, which also is associated with this person, such as the affiliation, e-mail address, and participation in conferences or organizations.

The subgraph for one person is shown in Figure 2. A phonetic mapping is created for each "pers" label, which also is linked with potential alternate-name spellings. Alternate spellings are generated based on names already seen, so no external input is required. Figure 2 also shows the common phonetic representation of Andersen and Anderson. Both of these entries were found in the raw data. The algorithm does not decide which is right or wrong; it simply

DOI: 10.1002/asi

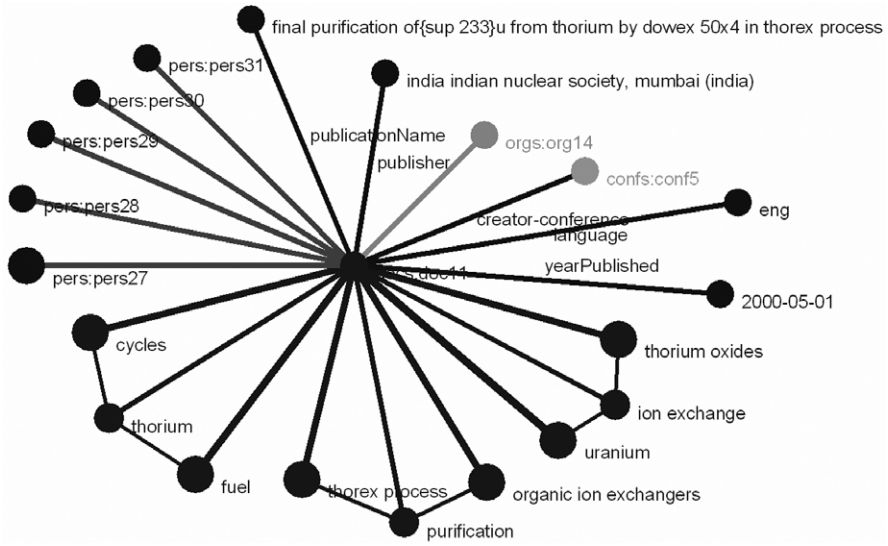


FIG. 1. Triple representation for one record (Document 11).

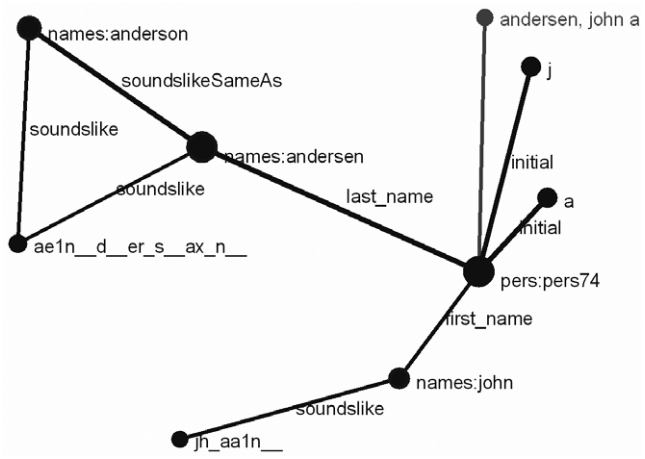


FIG. 2. Triples for one person (pers74).

establishes that they "sound" the same and that these two labels therefore are the same in the phonetic dimension. Note that this does not mean that they are references to the same person; that will be determined later.

Search Space Reduction

One major problem faced when disambiguating entities within a very large data set is knowing what entities to compare for disambiguation [since a brute force $O(n^2)$ comparison is most often infeasible]. In 2,000 bibliographic records, approximately 9,000 potentially unique individuals are identified. Beyond the obvious string matching, one must devise schemes for strategically comparing only those entities that are likely to be coreferent. In this case, the techniques used are phonetic representation and word distance. A sample graph is shown in Figure 3. By mapping names to their phonemes and finding similarly spelled names, the algorithm is able to identify authors who may be coreferent. This expansion allows it to capture both misspellings and alternate spellings.

A complete image, for 2,000 records, of the entire sound-alike (phonetic) and spell-alike (word distance) subgraph is shown in Appendix B.

Disambiguation of Author Entities

To determine the likelihood that two author labels reference the same person, a path search algorithm is used to calculate a weighted sum of the paths that connect them—comparing only those pairs that either have the same last name, similarly spelled last names, or same-sounding last names. Using this weighted sum, the "*owl:sameAs*" predicate is created when the likeness score is higher than a specified threshold. The use of this RDF predicate essentially collapses the two nodes into one, thereby generating a more compact graph (for a depiction of the searched paths, see Figure 4).

Note that "*owl:sameAs*" is an RDF predicate notation, which means that in the Web Ontology Language (OWL) domain, the Object and Subject nodes are to be considered to be the same once this triple is instantiated. From here forward, it will be referenced by using the simplified "*sameAs*" label.

High-level details of the current inferencing algorithm, shown in Figure 4, are as follows:

1. Person entities are considered for comparison if they share any of the following:
   a. The same complete "label." This is the string from the original record and may contain first name, last name, initials, name suffix, and name prefix.
   b. At least one identical last name

DOI: 10.1002/asi

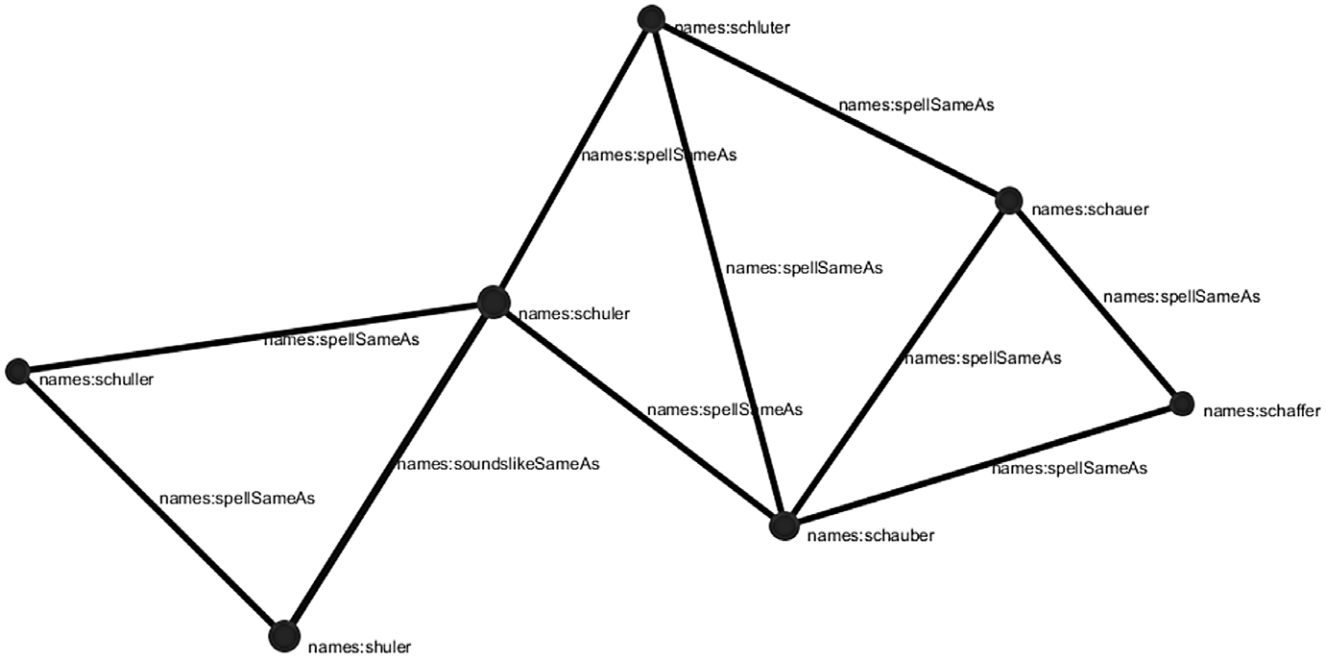


FIG. 3. Phonetic and word distance dimension "*sameAs*" strings.

c. At least one same-sounding last name (name mapped into a phonetic dimension)
d. At least one similarly spelled last name (The Levenshtein algorithm is used to calculate word distance for those whose last names start and end with the same letters.)

2. Candidate entities are deemed to be the same unless conflicting information is found. Currently, the only potentially conflicting information used is first name(s) and/or initials. For example, two candidate entities are checked for matching initials and first names. If a last name matches, and the initials/first name from one entity are a subset of those from the second entity, they are considered a strong candidate. However, if a difference is found, it means that there is conflicting information, and the entities may not be coreferent. These candidates are given a negative weight for this path. It may be overridden by strongly connected entities as determined by dimensions considered in the next step.
3. Coauthorship, affiliations, overlapping publication themes add weight to the total score used to determine coreference.

The sum of the weighted paths is compared to a threshold value, and those entity pairs scoring higher infer a new "*sameAs*" triple denoting the candidates to be coreferent. The semantic network paradigm by definition collapses these nodes, and subsequent operations will use the combined values. For example, if the first comparison of two candidates yields a match, their data values are now combined, and a third like candidate is essentially compared to the two merged records. Therefore, if one of the first records contains a full first name, it will be used for subsequent comparisons. As more nodes are collapsed, the set of known information for an author grows, but the criteria for candidate matches narrow in some dimensions. For example, first names may now be specific because they may have been fully spelled on a record as opposed to having initials only. However, the list of publication themes grows along with the list of known coauthors for that individual that could now have paths to other potential coreferences.

The prototype framework developed was used in the study documented in the Results section. As identified earlier, the framework uses several open-source algorithms for word distance, phonetic representation, and a commercial algorithm for theme extraction. Code developed at LANL makes the different algorithms work seamlessly and implements the weighting calculations that infer coreference. The inferencing algorithm generates additional "*sameAs*" triples over several passes until no new triples are added and the system stabilizes.

Results

A simple analysis of some 100 bibliographic records generated some 8,000 facts, thereby creating a semantic network with 8,000 triples (edges). One hundred documents is a very small set when considering real-world problems. It is envisioned that real-world problems would generate very large semantic networks and require massive computing resources such as those available in LANL's High Performance Computing (HPC) Division. This would allow the processing of real-world usable sets in real time. Currently, the dimensions of 2,000 documents can be processed on a

DOI: 10.1002/asi

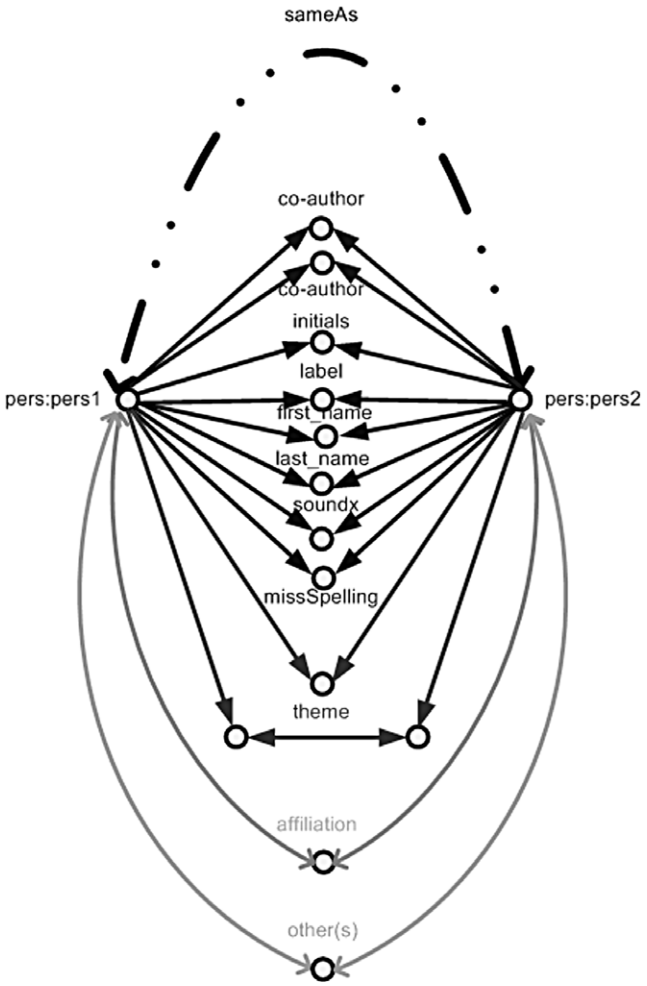


FIG. 4. Paths to be weighted between candidate entities to determine coreference.

capable workstation in 90 min. Note that the accuracy of these automated dimension extraction algorithms varies, and this in turn affects the accuracy of the disambiguation process. However, by combining the approach of the various algorithms, the number of false positives or missed coreferences is diminished, as will be shown later.

To ascertain the results, several tests were conducted using the framework. Some small sets were used to specifically test certain features. The small sets were extracted from larger sets and modified as needed to exercise selected parts of the framework. However, the final validation was done using the larger sets. For the results presented, three sets were used: a 50-record "small" set used to dissect a complete run; a 2,000 record "medium" set used for timing; and an approximately 4,400 record "large" set used to quantify the accuracy of the results.

Table 1 contains the first 25 unique persons identified for the small set. Originally, there were 142 person references extracted, and 23 of those were found to be coreferent, giving a list of 119 unique individuals. To infer coreference, a total of 33 candidate-pair comparisons were made.

For the entry "Ahn," there are two noncoreferent entities because the set of initials is different. Note that for some entries, a full spelling of the first name is listed whereas others only have initials. This depends on the raw data. The algorithm will use what it has to make an inference. Note that in principle, last name and initials matching is not sufficient to identify two authors as being coreferent; additional supporting information should be found. Differences in first names of candidate pairs generate a negative weight for that path which may be overridden when weights of other paths are added to the total score. This allows for some variations in coreferent entities' first names.

Figure 5 shows a subset of the results for the small case. The "pers:" prefix nodes represent person references, connected to the labels as found in original records. Then there are the inferred "*sameAs*" triples. The size of the nodes denotes the frequency of that label. For example, note the size of the "ahn, joonhong" node. This denotes that most of the references in this set use that label. Also note the variations of names and composite names matched.

The subgraph of all that is known for "Ahn, Joonhong" is shown in Figure 6. This subgraph shows information such as the papers published, the conference associated with the different papers, coauthors, themes, and more. This same output then can be formatted for human consumption, as shown in Table 2. These images are meant to illustrate that the graphs are very complex and use color to denote the dimensions.

In addition, "test" records were introduced or modified. One such test case was for the author "Anderson." One of the records was modified to have an alternate spelling. The algorithm still found the records to be coreferent based on content and shared coauthors. However, if the record was modified enough (e.g., different initials), no matches were made.

Other subgraphs also can be used to show selected dimensions such as the conceptual dimension shown in Figure 7. This subgraph is generated by selecting the themes from each of the abstracts. Themes are extracted automatically in the form of primary-theme → related-subTheme. Thematic relationships are not your traditional categorization, such as Biology having a subTheme of Microbiology. Instead, these relationships are drawn from the proximity of themes in the original text. In other words, one theme is found often in the proximity of another theme in the analyzed abstracts. Going back to Figure 1, the main themes for Document 11 are "thorium," "purification," and "ion exchange." For "thorium," the subThemes are "cycle" and "fuel;" for "purification," they are "thorex process" and "organic ion exchanges;" and for "ion exchange," they are "uranium" and "thorium oxides." Each of these theme → subTheme edges are incorporated in the graph shown in Figure 7. Nodes with the same label allow connections of themes from different documents, thereby generating a connected conceptual map. In this particular instance, the predominant themes are Uranium, Nuclear, Reactor, and Fuel. All these themes and their relationships were derived

DOI: 10.1002/asi

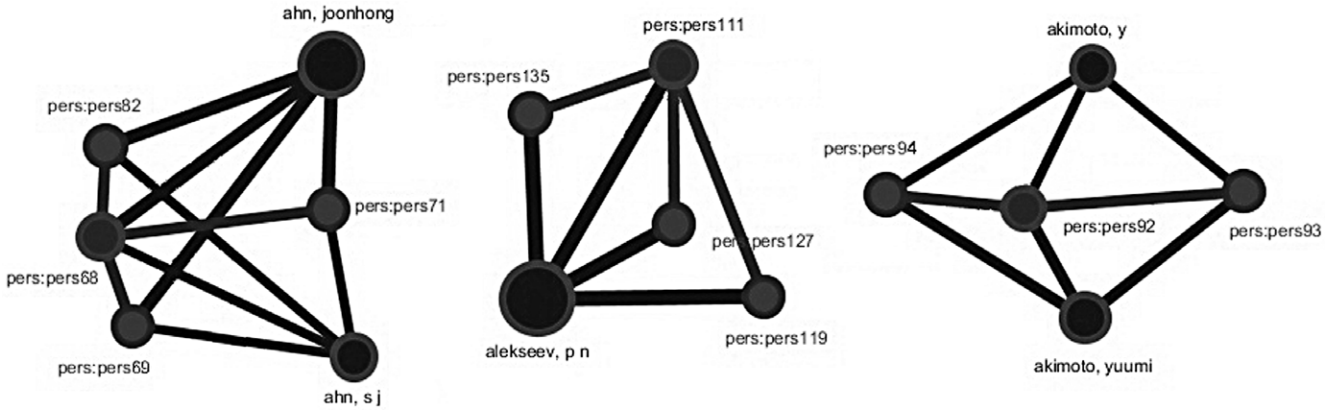


FIG. 5. Sample inferred *sameAs* entities subgraph.

automatically from the text of the abstract without human intervention. This automation allows the use of the framework in any context. The conceptual dimension subgraph is used in author disambiguation to find paths between the themes in the paper for one author and the themes in the paper of another author. The length of the path in the conceptual dimension is a tuning parameter for the disambiguation algorithm. This gives the ability to link entities that have been published on "related" topics, and does not require an identical match of themes for the candidate papers. As shown, the conceptual network quickly gets busy. The image in Figure 7 is meant to illustrate the complexity of the network, and an interactive interface should be used for detailed exploration.

Different tuning parameters were tested using the "medium" set. Results for variations in these parameters are documented in Table 3. There were a total of 5,904 person-associated entities extracted from the original records.

The table shows the weights used in the path for initials and the length of the "conceptual" path. A theme path length of "2" means that both documents must contain the same theme; "4" means that the there can be up to four connected themes in the "conceptual" path. On the top of Figure 7, we can see such a path starting in "thermodynamics properties" → "cerium" → "thorium oxides" → "ion exchange." For any path of length "X," the starting theme must be in one abstract, and the ending theme must be in the other. This allows "conceptual" proximity to be used as a weighted path when matching two entities.

In the table, "Trivial Label matches" are those where the entire label of an author matched the other entry. "Last Name match" means that one of the last names matched one of the last names of the other entity and that these "matches" therefore were candidates for further comparison. "Spell-alike" means that the last name of the candidate entries were different, but had similar enough spellings to be compared. The three match columns document the number of matches made based on the noted criteria. The next column is the total number of inferenced coreferences. The sum of the weighted paths had to be greater than or equal to 1.0 to infer coreference. This set did not have any sound-alike comparisons.

Note that when "Weight of Initials = 1.0," which means that first-name initials are enough to make a match, in this particular set, more entities were found to be coreferent. This leads to less compute time, as entities, once matched, do not have to be considered again in the future. In a small closed set, this optimistic match may work well; however, with a broader set using "initials," matching will contain false positives and therefore is not recommended in general.

For computing purposes, a path that is not longer than four themes seems to be a reasonable choice, as a very large set will have many paths between entities. All these paths will have to be evaluated, and this is an expensive graph operation. Also note that increasing the length of the theme path does not yield more matches in the current case. A longer path may pay off on rare occasions where the abstracts of two coreferent authors publish on marginally related conceptual areas, but it significantly adds to the compute time because searching for longer paths increases the compute time by an order of magnitude. The number of comparisons does not change significantly because most of the compute time is spent searching for longer paths for those entities that are not connected otherwise. The algorithm uses a depth-first path search function. Note that variations on the weights yield meaningful changes in the results, as noted in Table 3.

To calculate accuracy, the "large" case consisting of 4,440 bibliographic records with abstracts was used. This set contains 12,284 references to authors. There were 100,421 potential resolution candidate pairs, and 4,002 high-confidence coreference inferences. Of these high-confidence inferences, 2,392 were Trivial Label (full-name) matchings, 1,565 were due to just Last Name matches, 19 were Sound-alike matches, and 26 were Spell-alike matches. This leaves a total of 8,282 (12,284 – 4,002) unique person references. Of the unique entries, 291 were references to people with the label "unknown" as specified in the original record, which were ignored in the analysis. The remaining nearly 8,000

DOI: 10.1002/asi

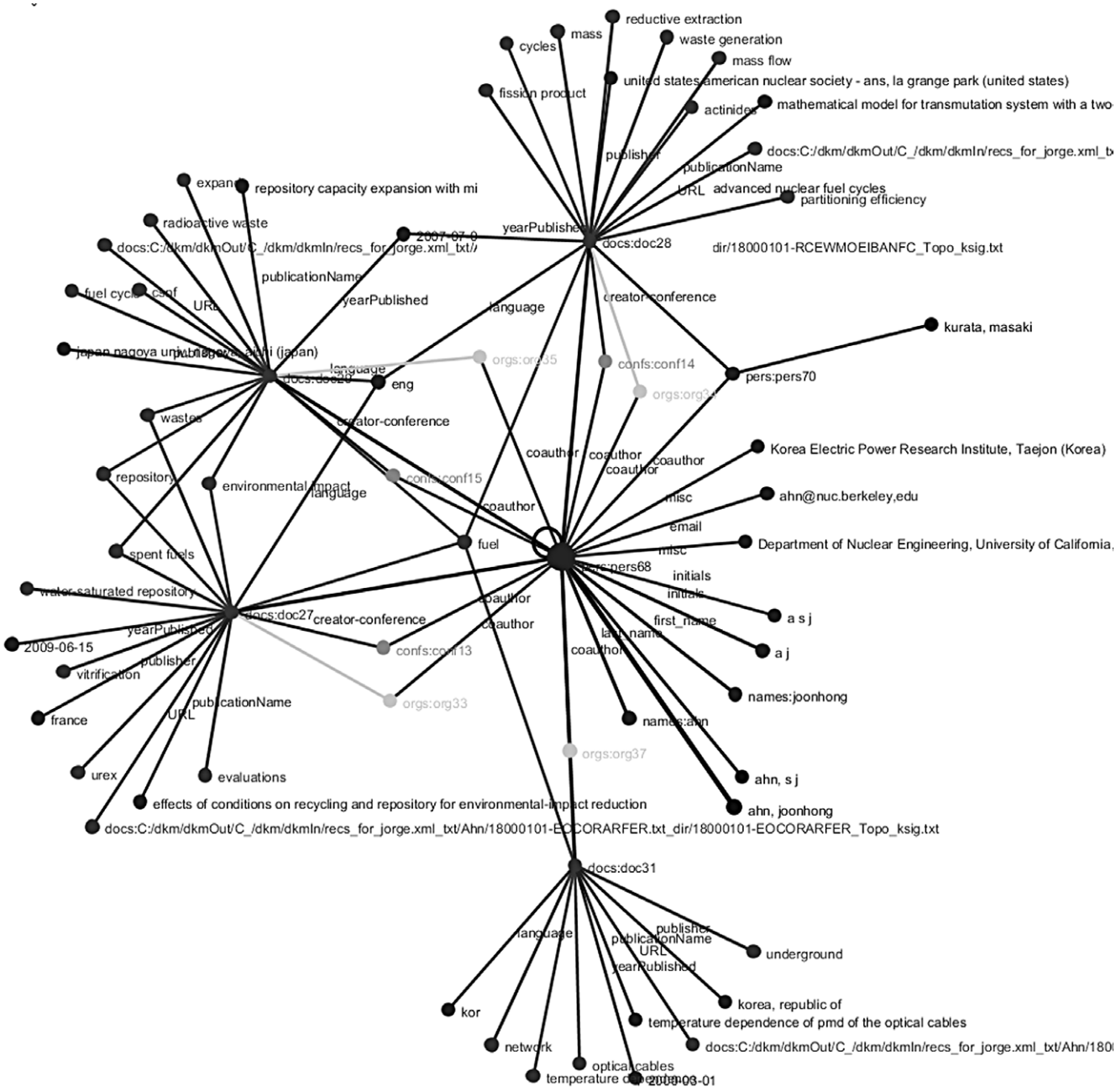


FIG. 6. Subgraph for author "Ahn, Joonhong."

unique references were checked by hand to see if any potentially missed matches could be identified. There were three types of potentially missed matches. In the first type, there were nine groups of same last name and same initials but no other paths, with a total of 18 unique entities in question. The second type contains first name or initials variations, with 19 groups containing a total of 40 entities. They had conceptual paths, but no coauthors, and therefore it was not enough to determine coreference given the name variations. The third type had last name variations and consisted of three groups with six entities. Together, these three types had a total of 64 potentially missed matches; this is less than 1%. However, note that most of the missed opportunities did not contain enough paths to determine coreference. Additional information would be needed to make a final determination.

Verification of a randomly selected 95 matches and missed opportunities entries yielded one false positive, two missed "conceptual" matches because the bibliographic records did not contain an abstract, and three missed last name variations because the entries have no shared coauthors. For the first two missed matches, a conceptual path was not possible. The remaining three missed matches did not contain shared coauthors, yet all these entries did seem to reference the same individual. One of these missed

DOI: 10.1002/asi

TABLE 2. Summarized facts compiled for "Ahn, Joonhong."

Last Name(s): Ahn
First Name(s): Joonhong, Joon, Hong
Initial(s): A J H, A J
Affiliation(s):<
Theme(s): actinides, biosphere, codes, csnf, cycles, environment, environmental impact, evaluations, expand, fission product, fuel, fuel cycle, iaea, mass, mass flow, nuclear fuel, nuclear fuel cycle, partitioning efficiency, partitioning process, radioactive waste, radionuclide, reductive extraction, repository, separation, separation processes, spent fuels, urex, vitrification, waste generation, wastes, water flow, water-saturated repository
Paper(s): Effects Of Conditions On Recycling And Repository For Environmental-Impact Reduction
Environmental Impact Of Nuclear Fuel Cycle And Application Of Compartment Models
Mathematical Model For Transmutation System With A Two-Member Chain And Variable Separation Coefficients
Repository Capacity Expansion With Minimization Of Environmental Impacts By Advanced Nuclear Fuel Cycles
Status Of Iaea Crp On Study Of Process-Losses In Separation Processes In Partitioning And Transmutation Systems In View Of Minimizing Long-Term Environmental Impacts
Coauthor(s): Bimova, K C; Bychkov, A; Inoue, T; Kawasaki, Daisuke; Kim, Chang Lak; Koch, L; Kormilitsyn, M; Kurata, Masaki; Nagarajan, K; Nawada, H; Ye, Y; Yoo, J H

matches is described in detail in Appendix A. On the other hand, the algorithms for word distance and phonetic dimensions did find 26 correctly matched entries with different spellings of the last names.

Shown in Table 4 is a list of the top authors of the "large" set that is based on frequency. All these matches were verified either by hand, simple graph, or simple bibliographic search methods using name and document title values. For this set, the simple graph check generated a subgraph of all that was known about the author. Then, using selective filtering of low-degree nodes, common paths between the records were exposed that verified the strong connectivity between the entities in question. These paths most often were conceptual and coauthor paths. Only one false positive was found; this erroneous match had strongly connected entities with last name variations. The authors worked in the same institution and published in the same field, plus their names were spelled alike.

The algorithm does multiple passes until no new "*sameAs*" triples are inferred. New "*sameAs*" triples may lead to new coreferences being found based on new shared coauthorship. Therefore, the algorithm repeats the process in "like" but unmatched entities until no new triples are added. After the initial pass, new entries could still be added in an incremental manner; however, any new inferences can lead to new shared coauthorship and therefore all unmatched candidate entities would have to be considered again.

The final semantic network can be queried for

a. "All that is known about an individual" (see example shown in Figure 6)
b. "What is known about a certain concept"
c. "Subgraph" of derived "*sameAs*" entities, and
d. Ad hoc queries.

The framework has a graphical interface to display results. Many of the figures shown are screen captures of queries to the framework. The images show how quickly the network grows to proportions that are hard for a human to fully comprehend. The interface is interactive, which is hard to show in static images because they tend to be very busy.

Future Work

The framework supports the insertion of external data sets such as ontologies. For example, the Library of Congress-derived classification of scientific terms would increase the accuracy of conceptual matches. Using these definitions may add references from the added traditional classification scheme that may not be present in the automatically extracted "theme." Similarly, other algorithms that deal with regional names spelling may be used to identify name variations, especially for foreign names. A known shortcoming of the current implementation also could be alleviated by providing as input first name variations of English names such as "Bob" and "Robert" and "Chuck" and "Charles."

The problem is not a traditional HPC problem, which tends to be numeric in nature. But it is a very large symbol-manipulation problem that should work well in an HPC environment. It is data-intensive because the extraction process is automated and quickly generates large amounts of data. Furthermore, algorithms to efficiently represent all the entities and to operate on the very large symbolic data set are needed. Future releases of the AllegroGraph software will be more suited for cluster computing. AllegroGraph V4.0 currently has an application using 300 billion triples. The focus of current effort has been on accuracy of the approach. The next step would be to optimize the algorithm and implement performance enhancements.

The framework is generic enough that it could be used to disambiguate other types of entities such as references to organizations (e.g., affiliations in bibliographic records). Different sets of paths and weights would be used, but the approach is the same; word distance, phonetic, and conceptual dimensions will continue to be used. Geographical dimensions would be added and used to play a key role. Other external data sets such as DBPedia (http://dbpedia.org) also could be easily incorporated to create additional paths not present in the raw input data.

Conclusion

The results support the hypothesis that connectedness identifies coreference. The challenge becomes the generation of a well-behaved semantic network, given that different algorithms were used to generate the edges. Careful weighting associated with the different connecting paths

DOI: 10.1002/asi

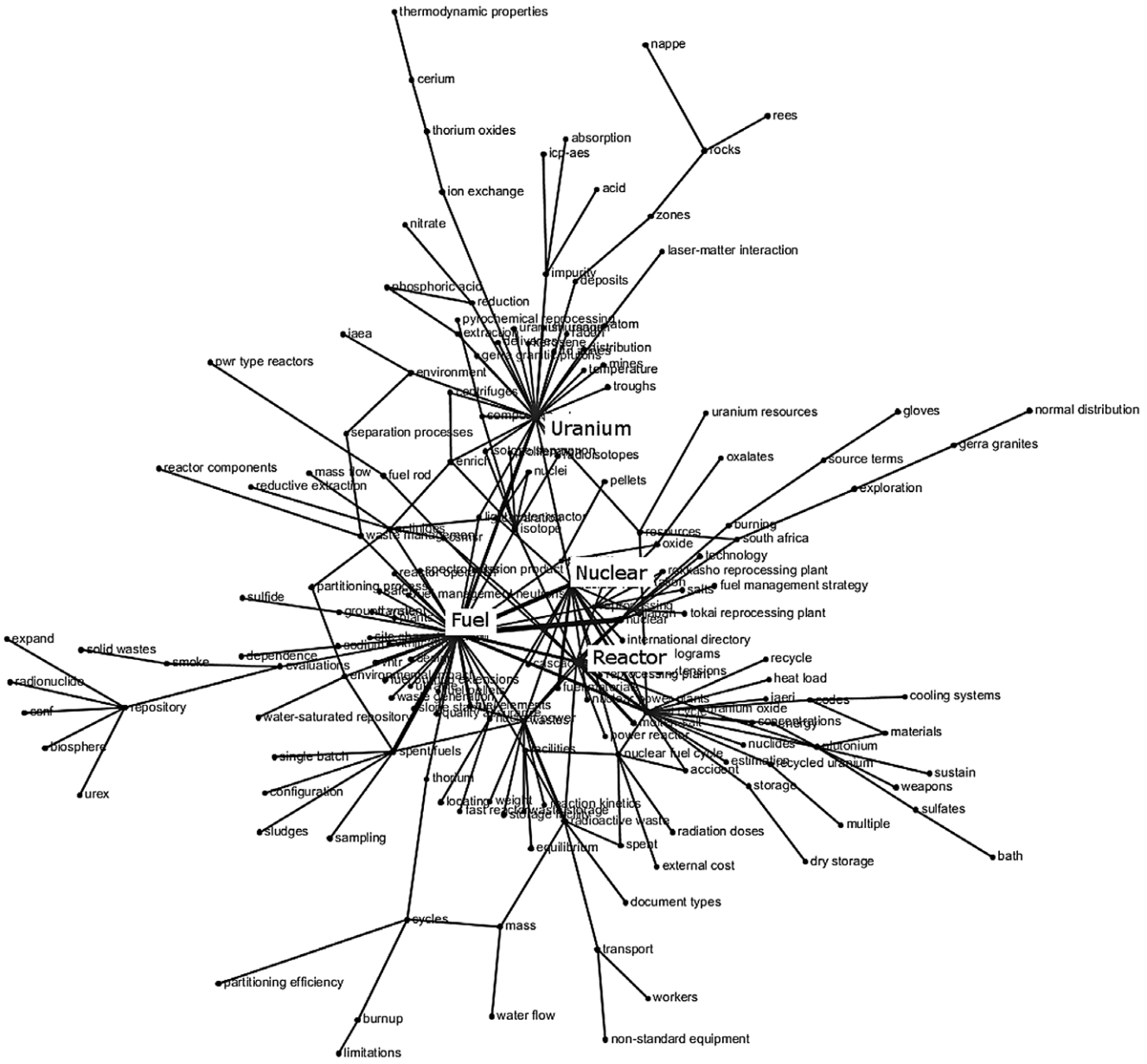


FIG. 7. Conceptual dimension subgraph.

TABLE 3. Comparative results of tuning parameters.

Weight of initials/Theme path length	Trivial Label match	Last Name match	Spell-alike match (details in Appendix A)	Total *sameAs*	No. of comparisons	Compute Time (min)
1.0 /2	844	21	0	865	6,621	59
0.91/2	440	17	0	457	9,539	93
0.91/3	593	20	1	614	13,749	101
0.91/4	604	19	1	624	13,685	298
0.91/6	604	19	1	624	13,684	2,769

also is important. The set of weights are an interpretation of what makes two entities coreferent since matches based on last name together with initials were deemed to be insufficient due to the potentially large number of false positives. The weights are crucial in reducing the number of false positives while minimizing the number of missed coreferences. In the current test set, the algorithm has a 99% accuracy rate based on the data provided. Note that the

DOI: 10.1002/asi

TABLE 4. Top-10 authors of the "large" set of 4,440 abstracts.

Author name(s) [Last, First]	Initials	No. of *sameAs* (coreferent entities)
inoue, tadashi	(i t)	24
van katwijk, ____	(v k c)	22
ko, won-il	(k w i a o)	21
park, seong-won	(p s w)	19
baron, pascal	(b p)	17
glatz, jean-paul	(g j p)	17
yang, myung	(y m s)	17
bychkov, ____	(b a v)	15
morita, yasuji	(m y)	14
cuney, ____	(c m)	14

bibliographic records are very good to begin with; however, different record sources introduce errors and duplication, and disambiguation therefore is still needed. Specifically, the test has shown that when no connectivity is found, no coreference is inferred, even when the names are the same. For establishing possible coreference in such cases, the algorithm would need additional information, just as would humans, to make a proper determination.

The ability to automatically fuse information from databases and unstructured text in one framework facilitates the analysis of very large sets with little human intervention. Coreference is just one of the many interesting patterns found in these large sets. The usefulness of these patterns increases as redundancy is eliminated.

Acknowledgments

We recognize those who have contributed to making this work possible. Bibliographic records from Thompson-Reuters and the Web of Science were used in the test cases. LANL's Research Library staff was instrumental in being able to generate interesting bibliographic sets for testing. Similarly, LANL's High Performance Computing Division support was crucial for the preparation of this publication and preliminary benchmarking on High Performance Clusters. In addition, Susan K. Heckethorn, Daniel A. Alcazar, and Matthew F. Hopkins from the Research Library were instrumental in their support and for helping with hand-validation of the matches. Hans Ziock provided excellent comments and a review from an outsider's perspective.

References

Bell, G.B., & Sethi, A. (2001). Matching records in a national medical patient index. Communications of the ACM, 44(9), 83–88.

Blanchon, H., & Boitet, C. (2006). Annotating documents by their intended meaning to make them self explaining: An essential progress for the semantic web. In H. Larsen, G. Pasi, D. Ortiz-Arroyo, T. Andreasen, & H. Christiansen (Eds.), Proceedings of the 7th International Conference on Flexible Query Answering (pp. 601–612). Berlin, Germany: Springer.

Collins, L.M., Román, J.H., Powell, J.E., Jr., Martinez, M.L.B., Mane, K.K., Xiang, Y., . . . and the Florida Department of Health. (2009). Using text analysis to reduce information overload in pandemic influenza planning. In Proceedings of the Sixth International Conference on Information Systems for Crisis Response and Management, Special Session on Solutions for Information Overload, Göteborg, Sweden (LA-UR 09-02971). Available at: http://www.iscram.org/ISCRAM2009/papers/

Durham, T. (1984). Fifth generation fever. Practical Computing, 7, 115–117.

Iosif, E. (2010). Unsupervised web name disambiguation using semantic similarity and single-pass clustering. In S. Konstantopoulos, S. Perantonis, V. Karkaletsis, C. Spyropoulos, & G. Vouros (Eds.), Proceedings of the Sixth Hellenic Conference on Artificial Intelligence (pp. 133–141). Berlin, Germany: Springer.

McRae-Spencer, D.M., & Shadbolt, N.R. (2006). AKTiveAuthor, a citation graph approach to name disambiguation. In Proceedings of the 2006 IEEE Computer Society/Association for Computing Machinery Sixth Joint Conference on Digital Libraries (pp. 53–54). New York: ACM Press.

Román, J.H., Collins, L.M., Mane, K.K., Martinez, M.L.B., Dunford, C.E., & Powell, J.E., Jr. (2008). Reducing information overload in emergencies by detecting themes in web content. In F. Fiedrich & B. Van de Walle (Eds.), Proceedings of the Fifth International Conference on Information Systems for Crisis Response and Management (LA-UR-08-2522) (pp. 101–107). Washington, DC. Available at: http://www.iscram.org/index.php?option=content&task=view&id=2236&Itemid=2

Román, J.H., & Spearing, A.S. (2009a, February). Knowledge fusion: Situation awareness through automated analysis of unstructured text. Presented at the 2009 Automated Fusion Integrated Capabilities Development Team (ICDT) Conference, Sierra Vista, AZ.

Román, J.H., & Spearing, A.S. (2009b, June). Discovery of patterns in digital records. Paper presented at the DESI III at ICAIL 2009, Global E-Discovery/E-Disclosure Workshop: A Pre-Conference Workshop at the 12th International Conference on Artificial Intelligence and Law, Barcelona, Spain (LA-UR-09-02958).

Román, J.H., & Spearing, A.S. (2009c, June). Derivation of knowledge from digital content. Presented at the 2009 Semantic Technology Conference, San José, CA (LA-UR-09-03135).

Rowe, M. (2009). Applying semantic social graphs to disambiguate identity references. In L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvönen, R. Mizoguchi, E. Oren, M. Sabou, & E. Simperl (Eds.), Proceedings of the Sixth European Semantic Web Conference (pp. 461–475). Berlin, Germany: Springer.

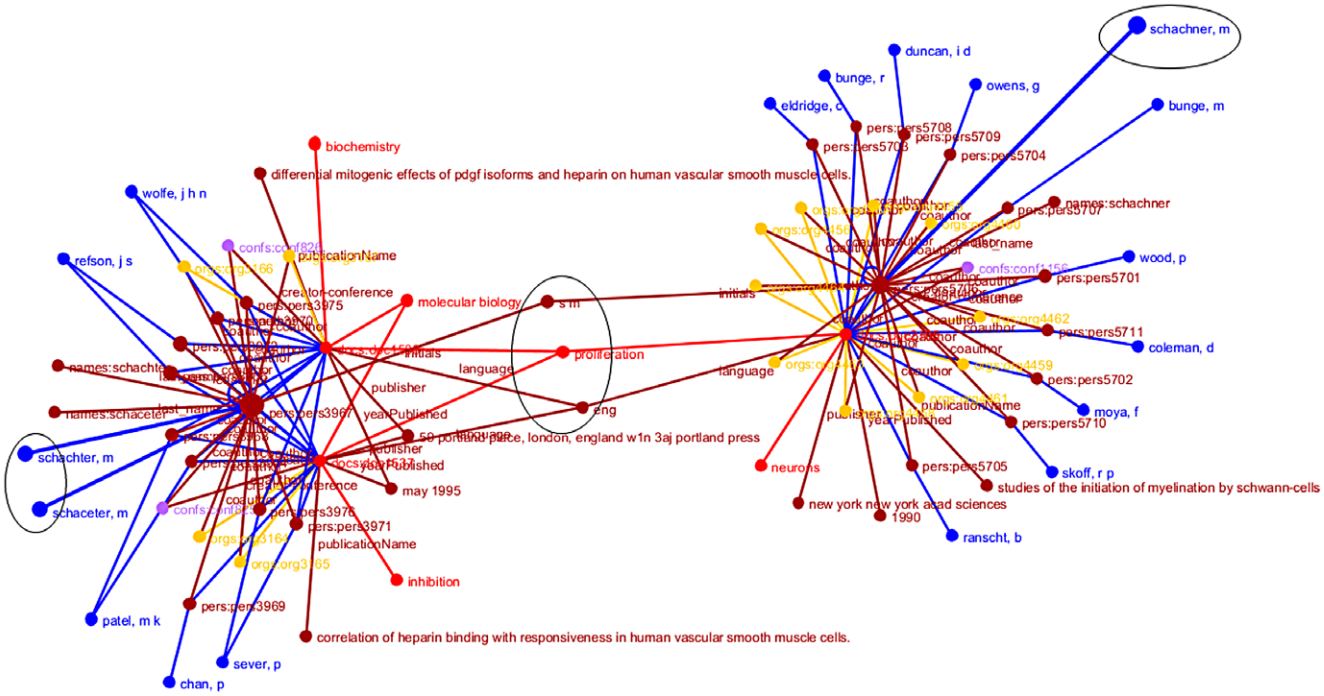
Turney, P. (1999, February). Learning to extract keyphrases from text. (National Research Council Canada Report Nos. ERB-1057, NRC-41622). NRC Institute for Information Technology.

Yerva, S.R., Miklos, Z., & Aberer, K. (2010). Towards better entity resolution techniques for web document collections. In Proceedings of the 26th International Conference of the IEEE Computer Society on Data Engineering Workshops (pp. 209–214). Piscataway, NJ: IEEE.

DOI: 10.1002/asi

Appendix A

Name-Spelling Variation Comparison



[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

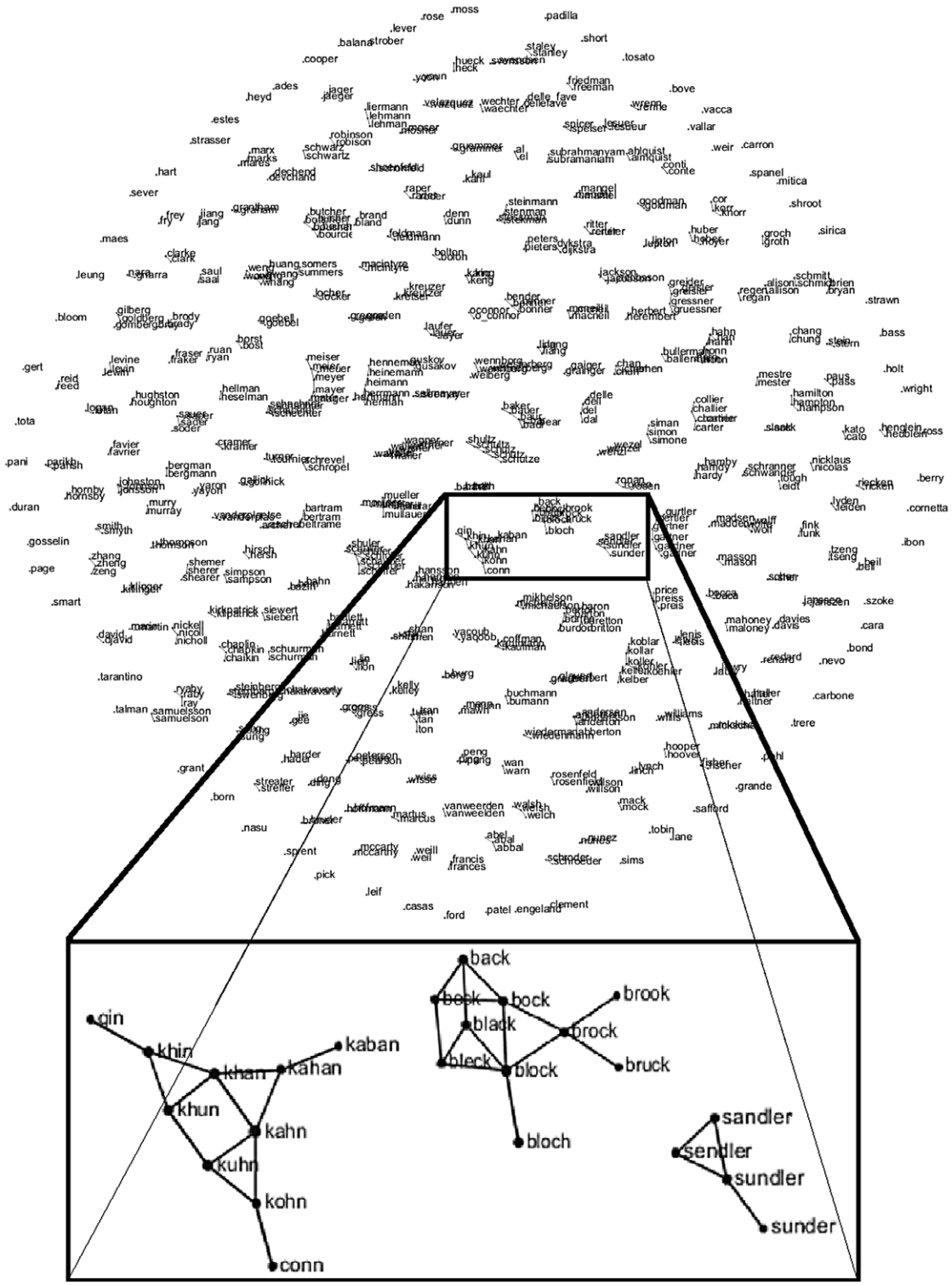
This subgraph shows the triples for three persons. Two of them, on the left side of the image, are inferred to be the same, and the third is not. The comparison of these persons is based on a word distance spelling algorithm for their last name. Person entities 3967, 3973, and 5706, with last names of Schaceter, Schachter, and Schachner, respectively, were compared for coreference. All three have the same initials: "S M." Note the three shared paths between the two sides of the graph; one is the "initials" triple with value "s m," the other is the conceptual path "proliferation" and the language fact "eng" for English language documents. The person on the right (Schachner) is not linked further to the other two persons and therefore is not deemed to be coreferent. In contrast, the two persons in the left (Schaceter and Schachter) are strongly connected, and therefore the weight of paths between them overcomes the initial lower score given to differently spelled last names. The documents corresponding to the two persons on the left also have five coauthors in common and share another theme of "molecular biology." The connectivity between these two persons is strong, as each of the five common coauthors started as separate entities and were found to be coreferent.

References to people use blue edges and nodes, themes are shown in red, general facts in brown, organizations in orange, and conferences in purple. Note that organizations and conferences are given generic names; these also would be candidates for disambiguation.

DOI: 10.1002/asi

Appendix B

Entire Sound-Alike (Phonetic) and Spell-Alike (Word Distance) Subgraph



This image is meant to illustrate that the subgraph for the phonetic and word distance dimensions is not totally connected. As shown, clusters of names identify the potential candidates for disambiguation based on these dimensions.

DOI: 10.1002/asi