Towards web supported identification of top affiliations from scholarly papers

David Aumueller

University of Leipzig
david@informatik.uni-leipzig.de

Abstract: Frequent successful publications by specific institutions are indicators for identifying outstanding centres of research. This institution data are present in scholarly papers as the authors' affilations – often in very heterogeneous variants for the same institution across publications. Thus, matching is needed to identify the denoted real world institutions and locations. We introduce an approximate string metric that handles acronyms and abbreviations. Our URL overlap similarity measure is based on comparing the result sets of web searches. Evaluations on affiliation strings of a conference prove better results than soft tf/idf, trigram, and levenshtein. Incorporating the aligned affiliations we present top institutions and countries for the last 10 years of SIGMOD.

1 Introduction

What are the most important research institutions in a specific field? Where do publications to conferences and journals come from? We want to answer such questions by analysing scholarly papers. Affiliations are stated in various forms denoting the same real world institution. Thus, to be able to aggregate papers by single real word entities, the different variants need to be matched. Affiliation variants include acronyms, abbreviations and multiple long forms, making it hard for common approximate string metrics such as edit-distance and trigram. We propose a web-based approach and **introduce the URL overlap similarity metric**. Each string is queried against a web search engine to collect the result set containing URLs to relevant pages. Basically we argue, the more URLs overlap across two result sets, the more likely the two query strings can be treated as synonyms. We assign countries to identified affiliations to depict their locations on a map, e.g. with more publications highlighted in different colour (more in shades from blue to red as in figure 1).

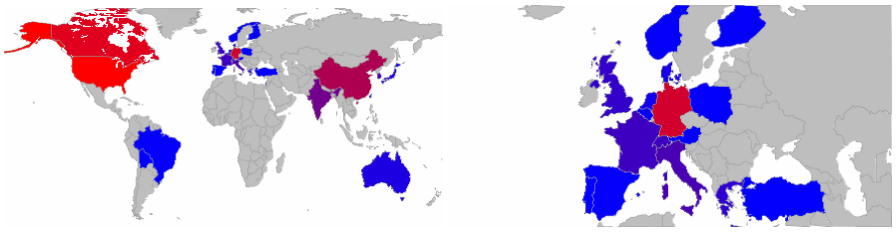

Figure 1: Maps of the world and Europe coloured by no. of 10 years' SIGMOD papers

Other motivating use of this data is the incorporation of an institution dimension as well as a geographic dimension in citation analyses, cf. [RT05], and online bibliographies such as collected using Caravela [AR07].

The next section outlines the general approach to identify outstanding institutions, in section 3 we present and evaluate our URL overlap metric for matching heterogeneous strings. Section 4 lists results of initial analyses of a conference series building on the data of section 3. We close by presenting related work and a summary with future work.

2 General approach to identify outstanding institutions

Authors of scholarly publications state their affiliations in the papers. As the ACM digital library website has these affiliations listed to each author we extracted the affiliation strings thereof (but are developing heuristics to extract such data from PDF fulltexts as well). The strings often contain departments and other details. We did no thorough pre-processing but left the strings as is except for cutting off the major variants of departments (suffixed by a colon) via a regular expression.

The next step is to determine affiliation strings denoting the same (real world) institution. We use our new approximate string measure (presented in the next section) to establish the mapping of corresponding affiliation strings. Groups of affiliation strings need to be clustered, e.g. by putting all strings into one group whose matches have a higher similarity score than a given threshold.

To analyse the data by geographical aspects, countries and possibly cities or latitude/longitude need to be linked to the institutions. We currently experiment with a large database of city names to establish such a mapping by matching city mentions. To disambiguate city names we iteratively decrease specificity: "*city, state, USA*", "*city, state*", "*city, country*", and "*city*" are used as patterns. Context in the fulltext of the papers may hint to determine the country as well, e.g. email addresses. For the results presented in section 4 we checked and completed the country information manually.

3 Approximate matching of affiliation string variants

Affiliation matching is a challenging task due to the many variants affiliations are stated not only on scholarly papers. The heterogeneity mostly derives from abbreviations of varying degree, ranging from ambiguous acronyms to self-explanatory abbreviations such as '*Univ.*' for '*University*'. One of the more challenging examples is the "*University of California at Santa Cruz, California*" of which the shortest form reads "*UCSC*", a medium short form "*UC Santa Cruz, Santa Cruz, CA*", and lots more are around. Common string metrics for matching two string variants include Levenshtein edit-distance, n-gram comparison, and soft tf/idf. For an overview and comparison of string metrics see e.g. [CRF03]. Applied to affiliations these work well within limits. These metrics do not help in mapping acronyms to long forms, i.e. they do miss a lot of correspondences such as stated above. In Wikipedia, mappings between varying

affiliation names are established for at least the well-known ones via numerous redirects pointing to the according article. One could extract and use this data, but scholarly papers and other sources contain more affiliations and variants than covered in Wikipedia.

3.1 The URL overlap string similarity metric

To overcome these drawbacks we propose a web-based approach that does not primarily take into account the overlap in the syntactic representation of the strings but compares the overlap of the search engine results queried for these strings. Here we concentrate especially on the overlap of identical URLs in the result set.

Hypothesis: The more URLs overlap the more similar the concepts behind the strings.

We use the complete URL for quantifying URL overlap and calculating similarity. We considered taking only parts of the URLs into account, e.g. domain or hostname similar to [Ta08], but web sites like Wikipedia are contained within the search result of many affiliation strings leading to identical hostnames that are not discriminating the search terms. Normalizing the URL overlap within the range 0..1 could be simply achieved by dividing the URL overlap by the maximum number φ of retrieved URLs, i.e. sim_1 = overlap / φ. We also take the ranks of the first overlapping hit into account and add that to sim_1 with a weighting factor (for examples of correspondences see table 1, next page):

$sim_{URLoverlap} = [\alpha \cdot (overlap / \varphi) + \beta / (1 + \delta)] / (\alpha + \beta)$, whereby …

overlap: number of overlapping URLs,
α, β: weighting factors, e.g. 2 and 1,
δ: distance between min ranks of overlapping URL, i.e. min(rank of URL u in result set for query str1) – min(rank of URL u in result set for query str2),
φ: number of retrieved search results per query, usually 10, 50, or 100.

We experimented with a URL overlap similarity metric that integrates also the ranks of all overlapping URLs with larger weights for higher ranks but could not detect more discriminating power.

3.2 Experiments and observations

To test our approach we collected over 4000 unique affiliation strings from database conferences covered in the ACM digital library. For each string we issue a web search query and store the maximum of returned results per one call into a relational database. With using the current BOSS service provided by Yahoo there is no restriction on the number of calls per timeslot. One call returns up to 50 hits, though. Each query result contains a projected number of overall hits and each hit entry contains (among others) rank, URL, size, date, and a snippet. To create the whole mapping of correspondences determined via URL overlap similarity a single SQL query suffices. A self-join on the web search result table as r and s using r.URL=s.URL aggregated by r.string and s.string

produces as group count the number of overlapping (identical) URLs. One could further specify a threshold in the having clause of the aggregation to limit the size of the resulting mapping. As the approach is relatively restrictive in nature we did not use a threshold here:

```
insert into mapping
select r.id, s.id, r.string, s.string, count(*) as overlap,
(2*overlap/50) + 1/1+abs(min(r.rank)-min(s.rank)))/3 as sim,
from result r, result s
where r.url=s.url and r.id<>.s.id
group by r.id, s.id
```

We experimented with levenshtein, trigram, soft tf/idf and URL overlap. We run soft tf/idf in the combination of a Jaro Winkler distance for the tokenized affiliation strings. The tf/idf-part takes care of scoring more frequent tokens lesser, which serves well for discriminating e.g. the lot of "*University of ...*" strings. Each set misses different potentially correct correspondences which would speak for a combination of similarity metrics.

URL overlap			
11, sim=0.48	Carnegie-Mellon University, Pittsburgh, Pa.	CMU, Pittsburgh, PA	true
10, sim=0.47	University of California, San Diego	UCSD	true
22, sim=0.46	University of Illinois, Urbana-Champaign, IL	UIUC, Urbana, IL	true
8, sim=0.44	Massachusetts Institute of Technology, Cambridge, Mass.	MIT, Cambridge, MA	true
5, sim=0.40	AT&T Labs-Research, New Jersey, NJ, USA	Bell Labs Research	true
6, sim=0.22	Dresden University of Technology, Dresden, Germany	TU Dresden	true
6, sim=0.11	Max-Planck Institute for Informatics, Saarbrücken, Germany	Saarland University, Saarbrücken, Germany	false
1, sim=0.02	University of Illinois at Chicago, Chicago, Illinois	University of Chicago	false
Levenshtein			
sim_l=0.69	Università di Torino, Torino, Italy	U. of Torino, Torino, Italy	true
sim_l=0.70	Michigan State University, Lansing, MI	Wichita State University, Wichita, KS	false
Trigram			
sim_t=0.50	Microsoft Corporation, Redmond, WA	Microsoft Research, Redmond, WA	true
sim_t=0.61	University of Illinois at Chicago	University of Chicago	false
Soft tf/idf			
sim_s=0.77	State Univ. of New York, Stony Brook	Stony Brook University	true
sim_s=0.37	Tsinghua University, Beijing, China	Microsoft Research Asia, Beijing, China	false

Table 1: Examples of true and false positives found by various sim metrics

The distance (difference) between the highest ranks of an overlapping URL supplies a further parameter in the score calculation, as hinted by the false[1] positive result of "*University of Illinois at Chicago, Chicago, Illinois*" ($rank_{highest}$=44) and "*University of Chicago*" ($rank_{highest}$=1), i.e. hit 44 for string 1 is the best whose URL is also in the result set for string 2, whereas already the URL of the first hit for string 2 is also within the result set for string 1. Considering the correct "TU Dresden" and the false "Saarland" exemplary correspondences of table 1, both have 6 URLs overlapping but the distances between minimum rank values are diametric and thus add more discriminating power to the measure, e.g. resulting in 0.22 for the correct match vs. 0.11 for the false match.

The choice of web search engine also influences the results [Sp06]. Some, e.g. Google, have synonym dictionaries in place that may prove useful here, but may be biased towards the English language. Also the available spell checkers and the resulting "did you mean"-links could be followed to sort out typos, e.g. the strings "*Univerisity*" and "*MicrosoftWay*" did not return useful results via the used Yahoo API, whereas in the end user interface such typos are corrected automatically. Besides, newly established institutions may be unknown or underrepresented in the web search results. Thus, no. of returned hits could be taken into account or maybe generally limited to a lower number than the used 50. Furthermore, the time querying a web service is costly and the web service acts as a blackbox that can only be adjusted within the available parameters.

Often, for manual labelling, the decision whether a match is a true or false positive is not an either/or-decision. Correspondences identified as false positives from a syntactic perspective could as well hint to an institution merger or name change as e.g. Bell Labs formerly known as AT&T. Generally, background knowledge is needed to decide whether two strings denote the same or different real world entities.

3.3 Evaluation

For evaluation purposes we will hand-proof precision and recall of a subset. For this we determine the true positives, i.e. correctly identified correspondences, as well as the false positives, i.e. false correspondences, of each tested similarity metric. Based on the cardinalities of these sets, we thus can compute precision and recall:

$$\text{precision} = |\text{true positives}| / (|\text{true positives}| + |\text{false positives}|)$$
$$\text{recall} = |\text{true positives}| / |\text{real correspondences}|$$

With precision the reliability of the found correspondences are judged whereas by recall the share of real correspondences that is found is nominated. To combine both measures in a single one f-measure can be calculated (see [DMR02] for a discussion):

$$F_\beta\text{-measure} = (1 + \beta^2) \cdot (\text{precision} \cdot \text{recall}) / (\beta^2 \cdot \text{precision} + \text{recall}), \text{ with } \beta \geq 0.$$

To quantify recall (and with precision also combined f-measure) the complete correct mapping is needed which we manually established for affiliation strings from the

[1] Not affiliated according to http://en.wikipedia.org/wiki/University_of_Chicago_(disambiguation)

publications of the SIGMOD 2007 conference as extracted from the ACM website. Here, 140 publications yield 150 different affiliations strings. We manually determined the perfect (symmetric) mapping having a size of 268 by determining clusters of affiliation strings and summing up the possible match correspondences. Only one third of the strings have correspondences, i.e. few institutions are represented by many affiliation variants whereas the majority of institutions only appear once. The following charts show lines for precision, recall, and f_1-measure (precision and recall evenly weighted).

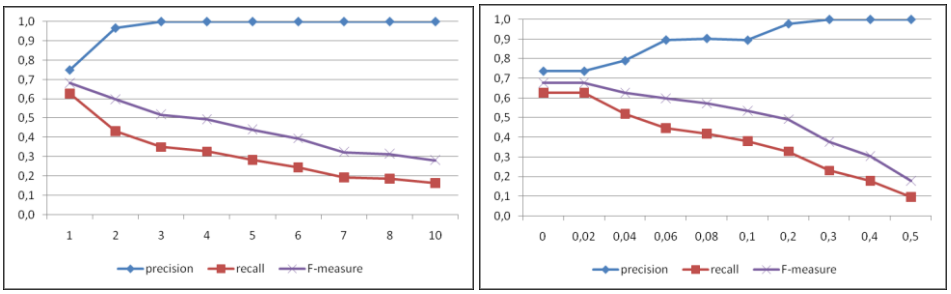

Figure 2: URL overlap similarity with at least a) x overlapping URLs and b) $sim_{URLoverlap}$

Concluding, according to this evaluation URL overlap (68% f-measure, figure 2) performs best for matching affiliation strings, followed by the soft tf/idf approach (55% at threshold 0.5, figure 3). We also tried trigram (47% f-measure) which still performed better than levenshtein (32%).

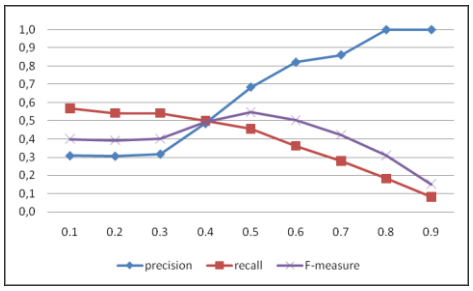

Figure 3: Soft tf/idf similarity with threshold t

Generally, the combination of URL overlap with classic string matching algorithms needs further testing and evaluations, as e.g. trigram or soft tf/idf combined with URL overlap to get acronyms/abbreviations could perform well in this context. Also left out in these experiments are manually established synonym tables for common abbreviations and blocking strategies as e.g. comparing academic affiliations and others separately. The danger here though lies in leaving out possible true correspondences, e.g. creating a subset of affiliations containing "Uni%" the "U of" matches would be lost. Further, more sophisticated pre-processing, e.g. not only separating off department but also institution name and location information, could boost similarity scores. Considering the fulltext of scholarly publications, institutions and locations could be derived from the authors' email addresses. Generally, to further evaluate the approach experiments with other data sets are needed, e.g. conferences and journals are variably named and referenced.

4 Results for a series of conferences

For identifying top institutions and top countries we examine ten years of SIGMOD publications as presented on the ACM web site. The collected data (table 2) consists of 1,026 papers (research, industrial, demo) with an average of 3.5 authors per paper across 200 different institutions in 1,044 different strings (993 distinct with department cut off).

year	papers	affils	insts	year	papers	affils	insts
1999	85	111	54	**2004**	118	145	64
2000	84	134	64	**2005**	116	146	78
2001	84	132	60	**2006**	99	140	79
2002	82	125	57	**2007**	140	150	79
2003	86	104	51	**2008**	132	173	90

Table 2: The examined data, ten years of SIGMOD

For further analysis we grouped affiliation strings with high similarity together, completed and corrected these clusters manually. In the ten years SIGMOD set there are e.g. 80 variants of IBM affiliations (with departments already cut off), in 2007 e.g., we have the following 10 specimens. For our analysis we denoted "IBM" as institution across all instances but assigned different countries accordingly.

```
IBM Almaden Research Center, San Jose, CA
IBM T.J. Watson Research Center, Hawthorne, NY
IBM T.J. Watson Research, Hawthorne, NY
IBM Silicon Valley Lab, San Jose, CA
IBM Toronto Lab, Toronto, Canada
IBM, Beijing, China
IBM, San Jose, CA
IBM Almaden Research Lab, San Jose, CA
IBM India Research Lab, New Delhi, India
IBM Toronto Lab, Markham, ON, Canada
```

Other years offer further variants, e.g.:

```
IBM, Markham, On, Canada
IBM Watson
IBM T.J. Watson Research Center, Yorktown Heights, NY
```

4.1 Publications by institution

We list top institutions in table 3 (next page). Apart from the Hong Kong University of Science and Technology (HKUST) and the National University of Singapore (NUS) the top 10 is dominated by US institutions. A further non-US candidate in this list of over 20 papers in 10 years of SIGMOD is the Indian Institute of Technology (IIT) Bombay. The top German institutions are the University of Munich with 11 positioned papers, the MPI with 7, the TU Dresden as well as the TU Munich with 6 each, followed by the University of the Saarland (5), RWTH Aachen, U Mannheim, U Marburg (4 each), HU Berlin, U Halle, U Konstanz, and U Leipzig (3 each).

	'99	'00	'01	'02	'03	'04	'05	'06	'07	'08	
IBM	6	11	11	19	11	15	14	10	18	20	**135**
Microsoft	7	6	8	5	8	15	14	10	19	11	**103**
Bell Labs+AT&T	14	11	10	16	12	13	5	8	5	6	**100**
Stanford Uni	5	7	5	2	10	6	4	1	2	1	**43**
U of Illinois	4	1	1	3	4	9	5	6	4	6	**43**
U of Wisconsin	6	3	1	6	2	10	3	5	4	1	**41**
NUS		3	2	1	4	3	9	3	3	11	**39**
UC Berkeley	1	3	3	6	8	4	2	4	3	1	**35**
Oracle	2	1	4	5		2	4	2	3	5	**28**
HKUST	2		1	3	3	2	3	3	5	5	**27**
U of Washington	2	2	4	2	2	5	6		2	2	**27**
Cornell Uni	3		2	4	3	3	1	3	5	2	**26**
U of Toronto			3	1	1	2	4	4	4	6	**25**
IIT Bombay	4	4	4	1	1	4	1	4		1	**24**
Carnegie Mellon		3	1	2	5	3	4	2	1	2	**23**
U of Michigan	1	1	1	2	2	4	1	4	4	2	**22**
U of Maryland	3	3	1	2	2	3		2	2	3	**21**

Table 3: Top institutions per papers in ten years (20+ papers)

4.2 Publications by country

For the following numbers (table 4) we have identified a country to each affiliation string variant, i.e. institutions may have multiple countries assigned. Breaking it down to states and cities is left to future analyses.

	'99	**'00**	**'01**	**'02**	**'03**	**'04**	**'05**	**'06**	**'07**	**'08**	**papers**	**insts**
US	65	64	68	64	69	92	89	79	98	84	**772**	584
CA	5	6	8	6	8	11	12	6	11	12	**85**	48
DE	8	8	4	6	3	6	7	8	13	7	**70**	67
CN	2		3	4	5	4	5	7	15	13	**58**	42
SG		3	2	1	4	3	9	5	4	12	**43**	19
IN	4	6	5	3	2	6	2	4	2	4	**38**	38
IT	2	4	2	1		3	4		6	3	**25**	31
KR	2	3	2	3	3			2	4	2	**21**	20
FR	2	4	2		2	3	1	2	2	2	**20**	27
CH	1		1	1	2	1	1	5	4	4	**20**	14

Table 4: Papers per year and country, as well as distinct institutions per country (countries with 20+ papers only)

In tables 4 and 5 countries are denoted by their ISO 3166 country code. Figure 1 on page 1 already illustrated these numbers visually on a map, as rendered by the Google charts API.

Having multiple authors from different institutions or countries on a single paper can be interpreted as co-operations. Table 5 lists the numbers of papers in the 10 year SIGMOD set authored by researchers from different institutions within the same or across two and more countries. All listed countries have papers co-authored with US institutions. Apart from these, only few countries have papers together with authors of other countries' institutions (notably exception being China and Singapore).

	US	GB	SG	NL	KR	IT	IN	IL	HK	FR	DE	CN	CH	CA	AU
AU	2													2	*3*
CA	44		4			2	5				3	3		*12*	
CH	6												*2*		
CN	18	2	12						2			*10*			
DE	22	2				2					*13*				
DK	4														
FI	2														
FR	10							2		*6*					
GR	6														
IL	3							*2*							
IN	21						*9*								
IT	9					*8*									
JP	6														
KR	12				*3*										
NL	2			*2*											
SG	14	2	*2*												
GB	9	*2*													
US	*285*														

Table 5: No. of co-operations between different institutions (2+ co-ops only)

5 Related work

In this paper we covered approximate string matching for heterogeneous variants. The linkage of short to long forms was studied recently in [Ta08]. The authors query web search engines with each form and link the short form (sf) to the long form (lf) if the lf is contained in the sf results' snippets and v.v. They also experimented with inverse hostname frequencies but did not take full URL overlaps into account, although the full URL is needed to discriminate e.g. the many Wikipedia hits. [El07] experimented with Jaccard similarity of hostnames returned by web searches, also neglecting the exact URLs. In the domain of bibliographic analysis the Citeseer project developed methods of extracting metadata from fulltexts, e.g. [Ha03] added initial support for affiliations. Location extracting or geotagging content is popular on the web, e.g. [Am04] disambiguates city names by taking context into account, [LB08] use unsupervised part-of-speech tagging to extract addresses, whereas we are concentrating on detecting location mentions by iteratively matching location strings of decreasing specificity. For erasing remaining ambiguities we plan to incorporate contextual information, e.g. email addresses. A recent citation analysis [RT05] also regards affiliations as extracted manually from the papers; for simplicity only the first author's institution were labeled,

though. As we collect data to all authors more types of analyses are possible, e.g. identifying possible co-operations between institutions.

6 Summary and future work

With the URL overlap similarity metric we presented a novel similarity metric for matching heterogeneous string variants denoting the same real world entity. The similarity builds on overlapping results of search engines queried with the strings. The affiliation match problem is a difficult one in that variants include acronyms and other abbreviations. We have shown that URL overlap outperforms levenshtein, trigram, and soft tf/idf in that task. Aligning affiliation strings are a needed step towards identifying outstanding institutions, i.e. only by clustering the variants to institutions and also locations (countries, states, cities) publications can be aggregated to project numbers of publications by institution and/or location. In this paper, we manually completed a first analysis of the last 10 years of SIGMOD publications. Results demonstrate again the dominance of US institutions (IBM with the most publications), followed by Canada and Germany as originating countries of many papers. With both the LMU and the TU, Munich can be seen as the top German city with the most publications (11+6) in this set.

We also see web applications making use of such data, e.g. the usual mapping mashup, probably with a timeline to illustrate not only the status quo but also change over time in the origins of publications to a conference series, journal, or topic. In web applications for categorizing publications, e.g. Caravela [AR07], institution and geographic location can serve as additional dimensions to search and navigate into the collection. Full automatic identification of outstanding institutions from scholarly papers is still to come, though. As related utilisation, names and locations of conferences and workshops could be aligned to categorize, rank, and map them by geographic location.

References

[Am04] Amitay, E. et al. Web-a-where: geotagging web content. SIGIR, 2004

[AR07] Aumueller, D., Rahm, E., Caravela: Semantic Content Management with Automatic Information Integration and Categorization. ESWC, 2007

[CRF03] Cohen, W., Ravikumar, P, Fienberg, S. A Comparison of String Metrics for Matching Names and Records. Data Cleaning and Object Consolidation, 2003

[DMR02] Do, H.H., Melnik, S., Rahm, E. Comparison of Schema Matching Evaluations. Web, Web-Services, and Database Systems, 2002

[El07] Elmacioglu, E. et al. Web based linkage. Web information and data management, 2007

[Ha03] Han, H. et al. Automatic document metadata extraction using support vector machines. Digital Libraries, 2003

[LB08] Loos, B., Biemann, C. Supporting Web-based Address Extraction with Unsupervised Tagging. Data Analysis, Machine Learning and Applications, 2008

[RT05] Rahm, E., Thor, A. Citation analysis of database publications. SIGMOD Record, 2005

[Sp06] Spink, A. et al. A study of results overlap and uniqueness among major web search engines. Information Processing & Management, 2006

[Ta08] Tan, Y.F. et al. Efficient Web-Based Linkage of Short to Long Forms. WebDB, 2008