Applications of Approximate Word Matching in Information Retrieval

J. C. French, A. L. Powell, E. Schulman*

Department of Computer Science, University of Virginia

*National Radio Astronomy Observatory

Introduction

Approximate Word Matching

– refinement of approximate string matching

– finer control over types of differences allowed

Introduction (cont.)

- **Overview of associated work**
  - **Motivation and application area**
  - **General approach**
  - **Issues remaining**
- **Approximate word matching**
  - **Definition**
  - **Uses**
  - **Evaluation**

What are we trying to do?

- **Merge bibliographic records**
- **Search distributed collections**
- **Bibliometric studies**

ADS

%T An H I survey of high-velocity clouds in nearby disk galaxies
%A Schulman, Eric; Bregman, Joel N.; Roberts, Morton S.
%J The Astrophysical Journal, vol. 423, no. 1, p. 180-189

SIMBAD

%T An HI survey of high-velocity clouds in nearby disk galaxies
%A Schulman, E.; Bregman, J.N.; Roberts, M.S.
%J Astrophys. J., 423, 180-189 (1994)

What is the problem?

Messy data!

- **variants**
- **misspellings**
- **acronyms and abbreviations**
- **multiple languages**
  - **translation and transliteration problems**

makes it difficult to tell when two strings represent the same entity

Our solution?

We use a combination of

- approximate data transforms
- string matching techniques
- approximate word matching techniques

to cluster the data and generate equivalence classes for entity names

Approach

- **Extract strings**
- **Cluster strings using chosen approach**
- **Domain expert reviews outcome of above; iterate until list finalized**
- **Utilize canonical forms and equivalence classes**

A running example

Astrophysics Data System (ADS)

- collection of bibliographic data, abstracts, and full text from astronomy and astrophysics
- approximately 240,000 entries from over 1,000 journals and conference proceedings
- our experiments are based on a subset of 146,000 journal articles

Data Set

- **120 affiliation strings representing 57 author affiliations in the states of Virginia and West Virginia**
- **57 equivalence classes**

Edit Distance

The edit distance, *e(u,v)*, from a string *u* to a string *v* is the minimum number of simple edit operations (insert, delete, replace, transpose) required to transform one string to the other.

Example,

e("Virginia", "Vermont") = 5

```
Virginia
Verginia
Verminia
Vermonia
Vermonta
Vermont
```

Raw affiliation strings for the University of Virginia

Affiliation string	Count
Univ. of Virgina, Charlottesville, VA, US	1
Univ. of Virginia, Charlottesvill, VA, US	1
Univ. of Virginia, Charlottesville, VA, US	44
Univ. of Virginia, Charlottsville, VA, US	1
Univ. of Virginia, VA, US	1
University of VA., Charlottesville	1
University of Virginia, Charlottesville, VA, US	23
University of Virginia, Virginia, US	1
Virgina Univ., Charlottesville, VA, US	1
Virgina, University, Charlottesville, VA	1
Virginia Univ.	2
Virginia Univ., Charlottesville	58
Virginia Univ., Charlottesville, VA	1
Virginia Univ., Charlottesville, VA, US	4
Virginia University, Charlottesville	1
Virginia University, Charlottesville, VA	1
Virginia, University	57
Virginia, University, Charlottesville	204
Virginia, University, Charlottesville, VA	77
Virginia, University, Charlottesville, Va.	83

564

Clustering Alternatives

- **Absolute edit distance**

$$e(u, v) \le \delta$$

- **Relative edit distance**

$$e(u, v) \le \alpha \min(u , v)$$

- **Approximate word matching**

Difficulties with Traditional Edit Distance

Moskovskii Gosudarstvenni Pedagogicheskii Institut, Moscow

Moskovskij Pedagogicheskij Gosudarstvennj University, Moscow

Edit distance 36

Alternative distance measures

Sorted surrogates

Original distance 36

`Gosudarstvenni Institut Moscow Moskovskii Pedagogicheskii`

`Gosudarstvennj Moscow Moskovskij Pedagogicheskij_University`

Distance 22

Approximate Word Matching

– **Given two strings *u* and *v*, find a minimum distance matching between the words in the strings.**
– **The sum of the edit distances is minimized.**
– **The cost of an unmatched word is the length of the word.**
– **Consider the sum of the edit distances.**

Moskovskii Gosudarstvenni Pedagogicheskii Institut, Moscow

Moskovskij Pedagogicheskij Gosudarstvennyj University, Moscow

Distance = 11

Approximate Word Matching vs. String Matching

```
s1: Moskovskii Gosudarstvennyi Pedagogicheskii Institut, Moscow
s2: Moskovskij Pedagogicheskij Gosudarstvennyj University, Moscow
s3: Virginia, University
s4: University of Virginia
s5: University of Vermont
```

$e(u, v)$

	s_1 (59)	s_2 (61)	s_3 (20)	s_4 (22)	s_5 (21)
s_1	0	36	50	50	50
s_2		0	45	52	51
s_3			0	17	16
s_4				0	5
s_5					0

$w(u, v)$

	s_1 (55)	s_2 (57)	s_3 (18)	s_4 (20)	s_5 (19)
s_1	0	11	56	52	52
s_2		0	48	44	44
s_3			0	2	7
s_4				0	5
s_5					0

Coincidences

```
Moskovskij Pedagogicheskij Gosudarstvennj University, Moscow
Virginia, University

Moskovskij Pedagogicheskij Gosudarstvennj University, Moscow
     V  ir     g  in    i       a,        University



Universitaetssternwarte, Vienna, Austria
Universitaet Sternwarte, Vienna, Austria
```

Evaluation Measures

– **Purity of Clusters =** of the clusters produced, the fraction that do not contain incorrectly placed items
– **Number incorrectly placed**
– **Number not placed**
– **Total misclassified =** number incorrectly placed + number not placed**.**

Clustering Experiments Using *e(u,v)* and *w(u,v)*

Distance measure	Relative distance (α)	Purity of clusters	Total mis-classified	Number misplaced	Number not placed
e(u,v)	0.20	78/79	29	1	28
	0.35	65/69	30	8	22
	0.50	55/62	30	13	17
w(u,v)	0.20	75/76	25	1	24
	0.35	62/66	23	4	19
	0.50	50/58	23	10	13

Finer Control

- **Some problems still remain**
  - **University of California, Davis**
  - **University of California, Irvine**
- **Constrain the allowable inter-word edit distance**
- **Use a Jaccard Coefficient to measure the degree of overlap**
- **Apply thresholds**

Clustering Experiments using a Jaccard Coefficient

Similarity coefficient	Purity of clusters	Total mis-classified	Number misplaced	Number not placed
0.75	78/78	24	0	24
0.65	68/69	15	1	14
0.50	56/59	11	5	6
0.40	48/54	13	10	3

Journal Title Clustering Experiments

Distance measure	Relative distance (α)	Purity of clusters	Total mis-classified	Number misplaced	Number not placed
e(u,v)	0.20	65/67	34	2	32
	0.35	46/54	31	10	21
	0.50	23/38	36	24	12
w(u,v)	0.20	60/61	29	1	28
	0.35	42/48	24	8	16
	0.50	25/38	31	23	8

Journal Title Clustering Experiments

Similarity coefficient	Purity of clusters	Total mis-classified	Number misplaced	Number not placed
0.75	70/71	39	2	37
0.65	49/60	39	12	27
0.50	29/43	37	23	14
0.40	29/42	36	23	13

Affiliation cluster/string	Number of occurrences
Virginia, University, Charlottesville	502
Virginia, University, Charlottesville	431
University of Virginia, Charlottesville	70
Virginia, University	59
University of Virginia	2
University of Virginia	1
University of Virginia, Virginia	1
University of VA., Charlottesville	1

Conclusions

- **Automated approaches can aid in the construction of equivalence classes.**
- **Approximate word matching is a useful tool for this activity.**

Acknowledgements

This work supported in part by:

- **NSF grant CDA-9529253**
- **DARPA contract N66001-97-C-8542**
- **DOE grant DE-FG05-95ER25254**
- **NASA GSRP fellowship NGT5-50062**

Approximate word matching

Moskovskii Gosudarstvenni Pedagogicheskii Institut, Moscow

Moskovskij Pedagogicheskij Gosudarstvennyj University, Moscow

Distance 11