

Information Extraction in the Large

CS 510 Winter 2007

1

What can you learn from the Web as a whole?

In particular, looking at co-occurrence as
relatedness

Not simply raw number of co-occurrences,
but deviation from chance

- Google Similarity Distance
- Lydia news analysis system

CS 510 Winter 2007

2



Google Similarity

Cilibrasi, Vitányi. *The Google Similarity Distance*, TKDE
March 2007

Determine a distance measure using
single and joint term frequencies on
Google (or other Web-search engine)

CS 510 Winter 2007

3



Information Distance

$$E(x, y) = K(x, y) - \min(K(x), K(y))$$

$K(\cdot)$ is Kolomogorov complexity – the length
of the shortest program to generate x
(or x, y)

Think of $K(x)$ as the compressed length of a
string x

This number is low if it is easy to get x from
 y (or y from x)

CS 510 Winter 2007

4



Normalized Information Distance

Normalize this value to run between
0 (strongly similar) and 1
(unrelated)

$$\frac{(K(x, y) - \min(K(x), K(y)))}{\max(K(x), K(y))}$$

CS 510 Winter 2007

5



Adapt to term frequency

Basically use inverse-document-frequency
weights

If x is a term, let \underline{x} be all documents
(on Google) containing x

$f(x) = |\underline{x}|$ number of documents with x

$f(x, y) = |\underline{x} \cap \underline{y}|$ num of docs with x and y

note $f(x) = f(x, x)$

CS 510 Winter 2007

6



Probability and weight

Probability of x,y : $g(x,y) = f(x,y)/N$

$$g(x) = g(x,x)$$

Here N is the sum of all $f(x)$'s and $f(x,y)$'s

Weight of x,y : $G(x,y) = \log(1/g(x,y))$

$$G(x) = G(x,x)$$

CS 510 Winter 2007

7



Normalized Google Distance

NGD(x,y) =

$$(G(x,y) - \min(G(x), G(y))) / \min(G(x), G(y))$$

Can put in terms of frequencies

$$(\max(\log f(x), \log f(y)) - \log f(x,y)) / (\log N - \min(\log f(x), \log f(y)))$$

- Ranges from 0 to ∞ (roughly)
 - Can go negative if Google reports $f(x,y) > f(x)$
- Choice of N doesn't seem critical, as long as it is larger than any $f(x)$
- Can use total number of documents

CS 510 Winter 2007

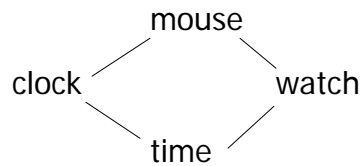
8



My experiment

I used $N = 100B$ (so $\log N = 11$)

<u>x</u>	<u>f(x)</u>	<u>log f(x)</u>	<u>x,y</u>	<u>f(x,y)</u>	<u>log f(x,y)</u>
mouse(m)	188M	8.27	w,m	28.1M	7.45
clock(c)	137M	8.13	c,m	7.1M	6.85
time(t)	2.12B	9.33	c,t	98.1M	7.99
watch(t)	294M	8.60	w,t	356M	8.55



CS 510 Winter 2007

9



NGD's

$$\text{NGD}(w,m) = (8.60 - 7.45)/(11 - 8.27) = .421$$

$$\text{NGD}(w,t) = (9.33 - 8.55)/(11 - 8.60) = .325$$

$$\text{NGD}(c,m) = (8.27 - 6.85)/(11 - 8.13) = .494$$

$$\text{NGD}(c,t) = (9.33 - 7.99)/(11 - 8.13) = .467$$

See experiments at

<http://clo.complearn.org/clo/listmonths/t.html>

CS 510 Winter 2007

10



What to do with NGD

- Clustering
 - Able to group paintings by artist, books by author
- Categorization by concept
 - Positive & negative examples
 - List of anchor terms a_1, a_2, \dots, a_n
 - Feature vector for term t
($\text{NGD}(t, a_1), \text{NGD}(t, a_2), \dots, \text{NGD}(t, a_n)$)
- Figure out word correspondences between languages