# **WEST**: WEIGHTED-EDGE BASED SIMILARITY MEASUREMENT TOOLS FOR WORD SEMANTICS

Liang Dong, Pradip K. Srimani, James Z. Wang
School of Computing, Clemson University

Web Intelligence 2010, September 1, 2010

# Outline

- Word Semantic Similarity and WordNet
- Concept Specification vs. Categorization
- Weighted-Edge Based Similarity Measure
- Experimental Studies
- WEST
- Conclusion and Future Studies

# Word Semantic Similarity (1)

- Word Sense Relation
  - Synonymy
    - “lumber” and “timber”
  - Polysemy
    - “bank”
      - Financial institution
      - Sloping mound

# Word Semantic Similarity (2)

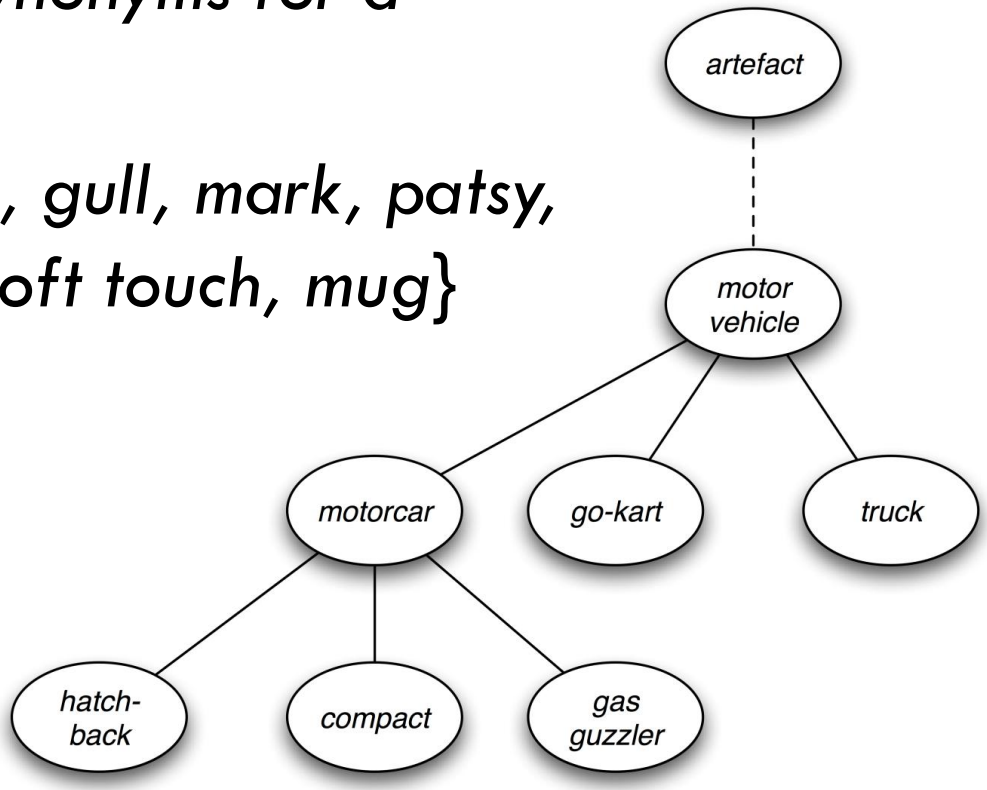
- Words Relations
  - Hypernym/Hyponym
    - "is-a" relation
      - Fruit-apple
      - mammal-dog
  - Meronym
    - "has-a" relation
      - A *wheel* is a part of a *car*
      - A *leg* is part of a *chair*

# WordNet 3.0

- WordNet lexical database
  - http://wordnet.princeton.edu/
- WordNet 3.0 release has 117,097 nouns, 11,488 verbs, 22,141 adjectives and 4,601 adverbs.
- The average noun has 1.23 senses and the average verb has 2.16 senses.

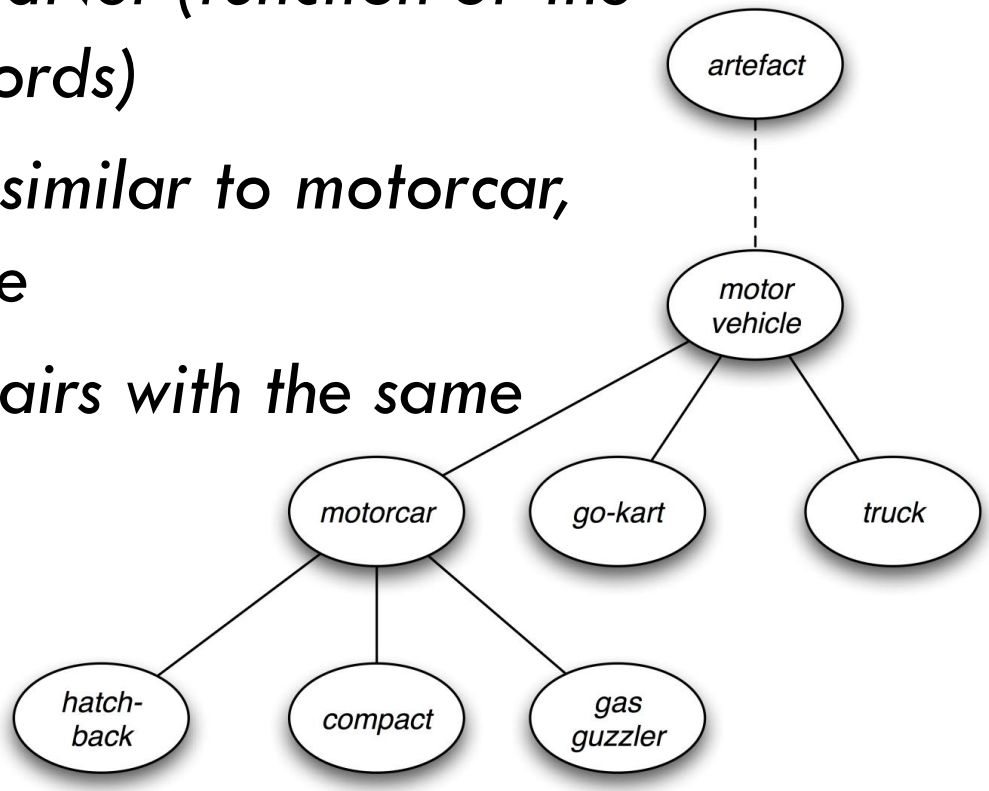
# WordNet 3.0

- *Synset: (Synonym set)*
  - *the set of near-synonyms for a WordNet sense*
  - *E.g. {chump, fool, gull, mark, patsy, fall guy, sucker, soft touch, mug}*



# Semantic Similarity of Words

- *One approach to determining the semantic similarity of words is by measuring their graph distance within the WordNet (function of the distance between the words)*
  - *Hatch-back is more similar to motorcar, than to motor vehicle*
  - *What about word pairs with the same graph distance*

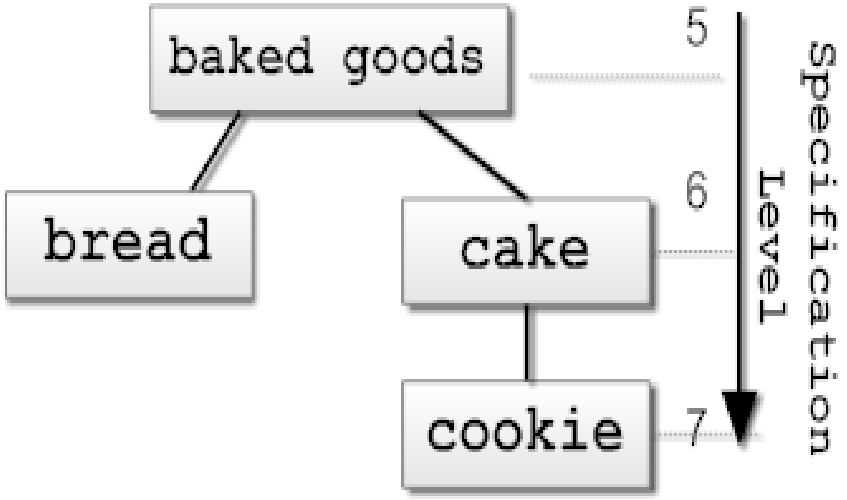


# Outline

- Word Semantic Similarity and WordNet
- Concept Specification vs. Categorization
- Weighted-Edge Based Similarity Measure
- Experimental Studies
- WEST
- Conclusion and Future Studies

# Specification & Categorization (1)

- Our recent study discovers that humans are more sensitive to the semantic difference caused by ***categorization*** than ***specification***.




| *Inheritance Word-Pair* | | *Categorization Word-Pair* |
|---|---|---|
| baked-foods :: cookie | ⇔ | bread :: cake |

# Specification vs. Categorization (2)

| *Inheritance Pair* | | | *Categorization Pair* | |
|---|---|---|---|---|
| baked goods :: cookie | 30 | ⇔ | bread :: cake | 19 |
| beef :: food | 48 | ⇔ | meat :: chocolate | 2 |
| brownie :: cake | 44 | ⇔ | cookie :: fruitcake | 5 |
| ground beef :: meat | 24 | ⇔ | pork :: mutton | 25 |
| apple pie :: pastry | 42 | ⇔ | pie :: puff | 8 |
| stove :: device | 41 | ⇔ | comb :: fan | 8 |
| engine :: machine | 18 | ⇔ | computer :: calculator | 33 |
| hunting dog:: canine | 27 | ⇔ | wolf :: fox | 22 |
| minicab :: car | 29 | ⇔ | jeep :: sedan | 21 |
| gold :: metal | 37 | ⇔ | aluminum :: zinc | 14 |
| **Total** | **340** | | | **157** |

**Groups with graph distance 2**

| *Inheritance Word-Pair* | | | *Categorization Word-Pair* | |
|---|---|---|---|---|
| apple pie :: food | 44 | ⇔ | cake :: beef | 3 |
| clementine :: fruit | 36 | ⇔ | apple :: almond | 15 |
| chicken :: food | 47 | ⇔ | octopus :: pastry | 0 |
| dynamo :: machine | 45 | ⇔ | engine :: abacus | 4 |
| abbey :: building | 26 | ⇔ | hostel :: mansion | 23 |
| tabloid :: medium | 8 | ⇔ | broadcasting :: journalism | 43 |
| laptop :: computer | 51 | ⇔ | workstation :: chatroom | 0 |
| American football :: athletic game | 36 | ⇔ | golf :: basketball | 14 |
| cliff diving :: sports | 44 | ⇔ | hunting :: swimming | 6 |
| collegiate dictionary :: book | 41 | ⇔ | atlas :: bestseller | 7 |
| **Total** | **378** | | | **115** |

**Groups with graph distance 4**

# Specification vs. Categorization (3)

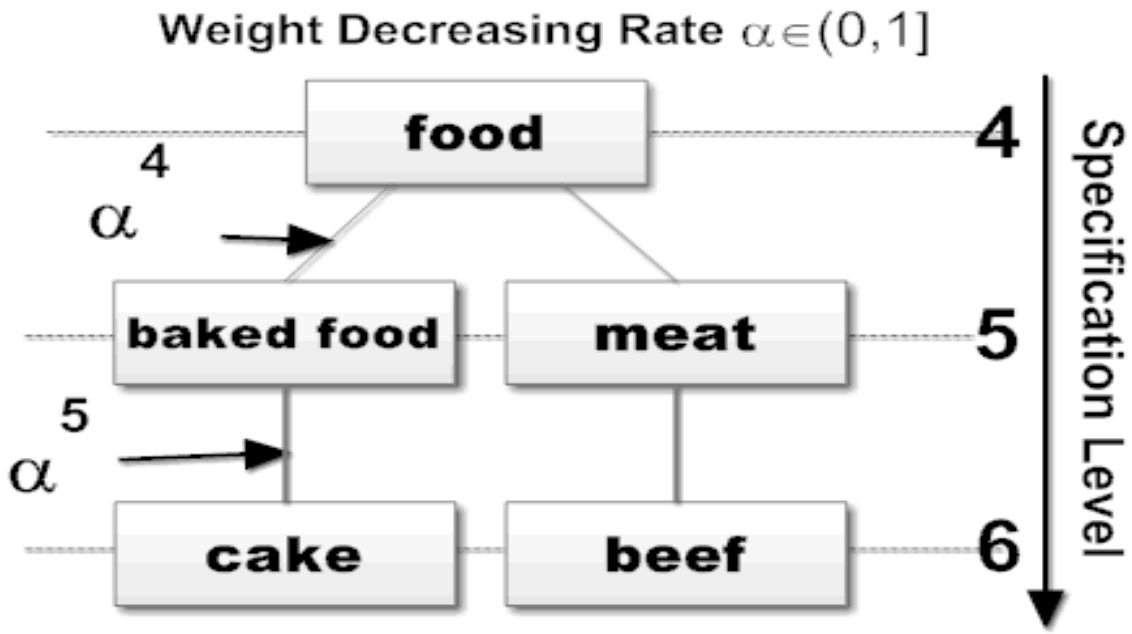
- we randomly stop people in Clemson University campus and ask them to judge which pair in each comparison group is more similar semantically. 51 individuals finished the questionnaire anonymously.
- The survey results in Table 1 show that in 68.41% of cases, people think the inheritance pairs are more similar, and in 31.59% of cases, people think the categorization pairs are more similar. The results in Table 2 demonstrate that in 76.67% of cases, people think that the inheritance pairs are more similar, and in 23.33% vice versa.

# Outline

- Word Semantic Similarity and WordNet
- Concept Specification vs. Categorization
- Weighted-Edge Based Similarity Measure
- Experimental Studies
- WEST
- Conclusion and Future Studies

# Weighted Edge Method

**Specification Level (SpecLev):** is the number of hops on the shortest path from the synset to its root, or the *depth* of synset in WordNet.



# Weighted Edge Distance Model

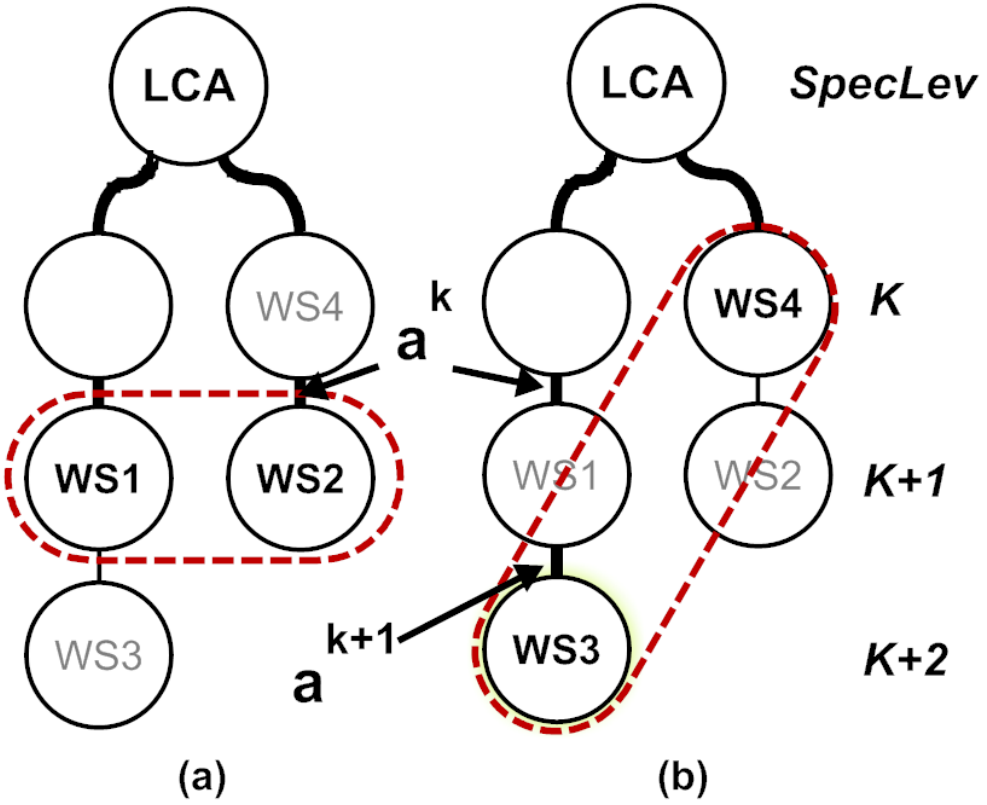
- Using our weighted edge model, we define *weighted edge distance* between a sense pair, as a function of the three SpecLev values.

$$l_w(ws_i, ws_j) = l_w(ws_i, ws_{root}) + l_w(ws_j, ws_{root}) - 2 \cdot l_w(ws_{lca}, ws_{root})$$

# Specification Level Difference

***Specification Level Difference* (SLD) :** absolute difference of the SpecLev of two word senses.

Increasing SLD from 0 to 2 reduces weighted edge distance




$$l_w(ws_i, ws_j) = \alpha^{slev_{lca}} \cdot \left( \sum_{m=0}^{slev_i - slev_{lca} - 1} \alpha^m + \sum_{n=0}^{slev_j - slev_{lca} - 1} \alpha^n \right)$$

# New Transfer Function g

- We define the semantic similarity between sense pair be a function of its weighted edge distance:

$$sim(ws_i, ws_j) = g(l_w)$$

- We use different non-linear functions and found that the similarity values obtained by hyperbolic functions best match human judgments.

# Key Features of WEST Approach

- Our weighted edge distance model uses the *SLD* count to implicitly differentiate the inheriting and categorizing relationships, matching the human perception.
- The new hyperbolic transfer function matches the human perception more accurately in transferring weighted edge distance into similarity value.

# Outline

- Word Semantic Similarity and WordNet
- Concept Specification vs. Categorization
- Weighted-Edge Based Similarity Measure
- Experimental Studies
- WEST
- Conclusion and Future Studies

# Benchmark Datasets

- Miller and Charles (MC) set as the comparison baseline. Since the earlier version of WordNet missed word "woodland" from the MC set, only 28 word pairs were used in these studies.
- We utilized all 65 pairs of the original Rubenstein-Goodenough set. Since MC set is a subset of RG set, we applied the 28 pairs of MC set as testing set $D0$, and the rest 37 pairs of words as training set $D1$.

# Strategy 1,3,5,6,7

$$sim_1(ws_i, ws_j) = g_1(l_{gd}) = e^{-\alpha \cdot l_{gd}}$$

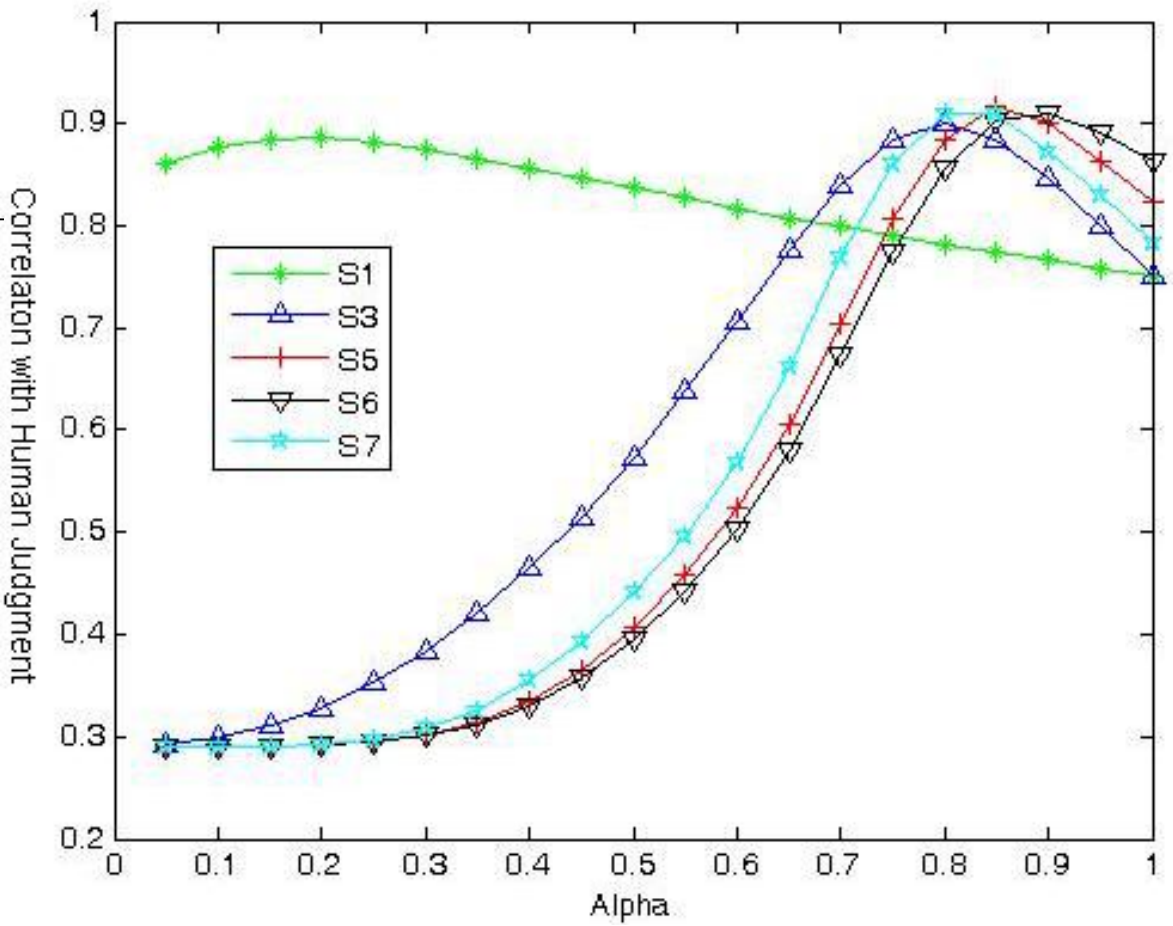
$$sim_3(ws_i, ws_j) = g_1(l_w) = e^{-l_w}$$

$$sim_5(ws_i, ws_j) = g_5(l_w) = \sec h(x) = \frac{2}{e^x + e^{-x}}$$

$$sim_6(ws_i, ws_j) = g_6(l_w)$$

$$= \tanh c(x) = \begin{cases} \dfrac{e^x - e^{-x}}{(e^x + e^{-x}) \cdot x}, & x \neq 0 \\ 1, & x = 0 \end{cases}$$

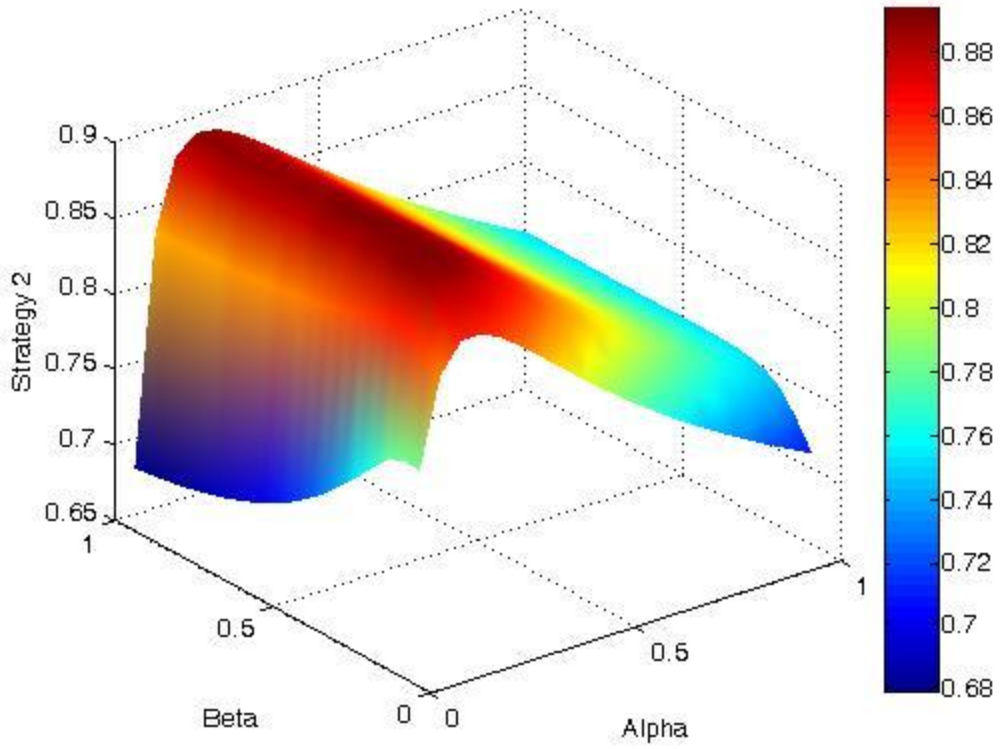
$$sim_7(ws_i, ws_j) = g_5(l_w) \cdot g_6(l_w)$$



# Strategy 2

$$sim_2(ws_i, ws_j) = g_1(l_{gd}) \cdot g_2(slev_{lca})$$

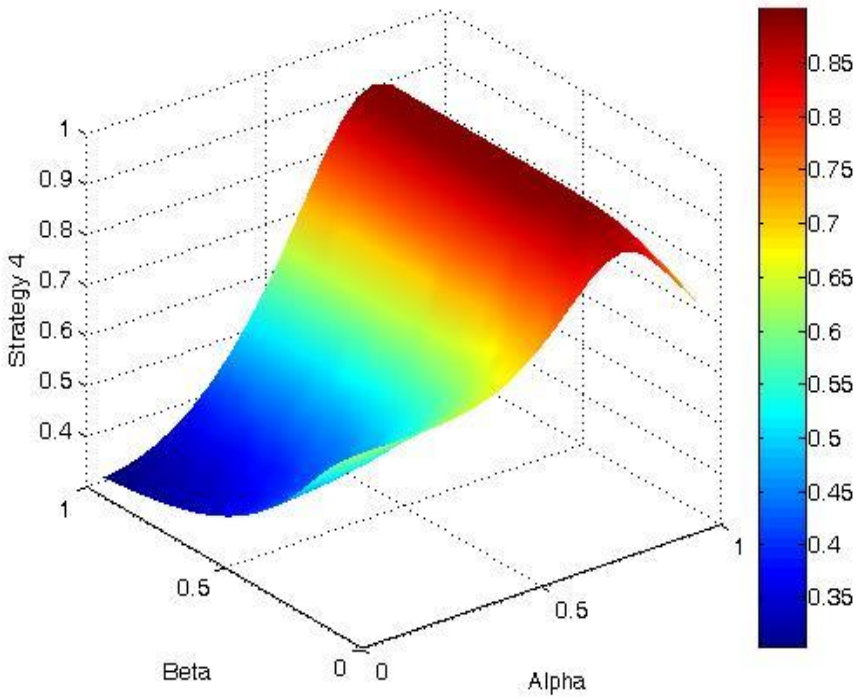
$$= e^{-\alpha \cdot l_{gd}} \cdot \frac{e^{\beta \cdot slev_{lca}} - e^{-\beta \cdot slev_{lca}}}{e^{\beta \cdot slev_{lca}} + e^{-\beta \cdot slev_{lca}}}$$



# Strategy 4

$$sim_4(ws_i, ws_j) = g_1(l_w) \cdot g_2(slev_{lca})$$

$$= e^{-l_w} \cdot \frac{e^{\beta \cdot slev_{lca}} - e^{-\beta \cdot slev_{lca}}}{e^{\beta \cdot slev_{lca}} + e^{-\beta \cdot slev_{lca}}}$$

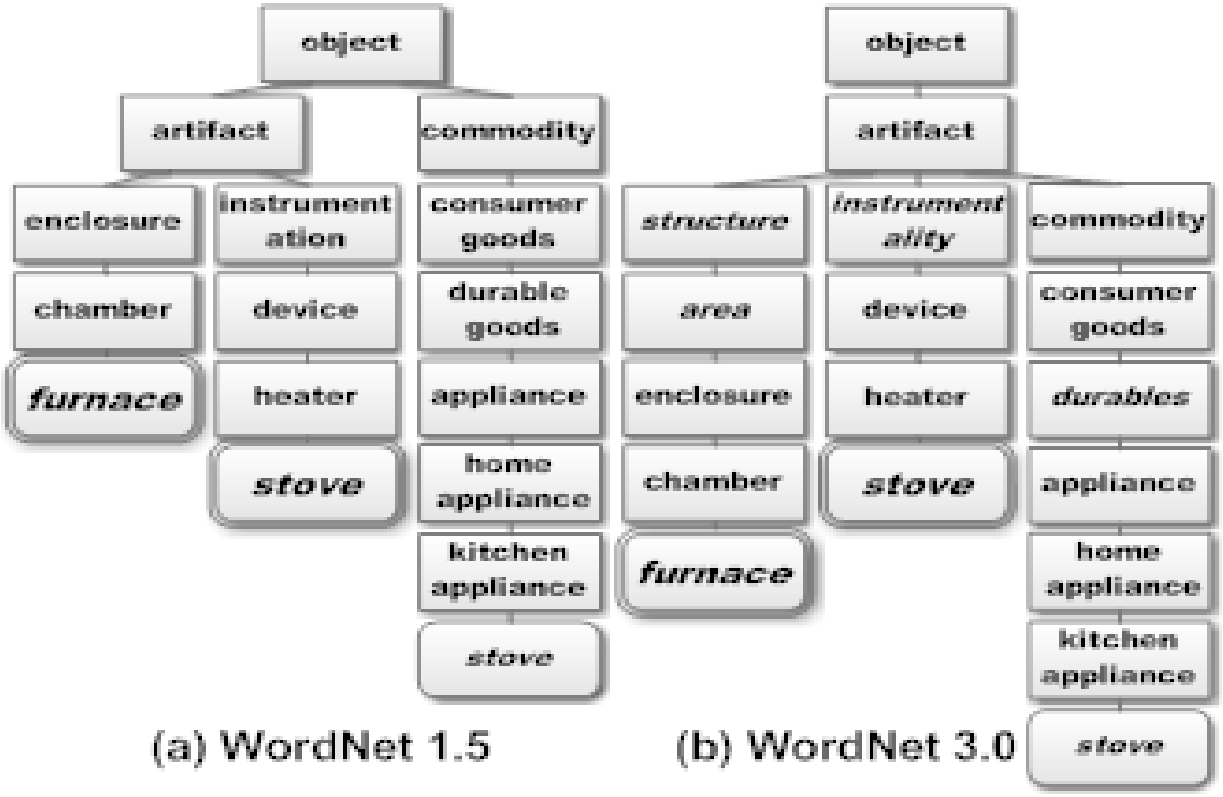


# Li's method with WordNet versions

The correlations between the computed similarities and human judgments using Li's method under different WordNet versions.

| | Li, 2003 | Varelas, 2005 | Us, 2009 |
|---|---|---|---|
| **WordNet** | 1.6 | 2.0/2.1 | 3.0 |
| **Correlation** | 0.8914 | 0.82 | 0.8078 |

# The Impact of WordNet Evolution



# Comparison with IC-approaches

| *Method* | *Type* | *G.Varelas, 2005* | *Us, 2009* |
|---|---|---|---|
| **Resnik** | IC | 0.79 | 0.8124 |
| **Lin** | Normalized IC | 0.82 | 0.7517 |
| **Jiang** | Hybrid | 0.83 | 0.6900 |
| **Li** | Graph Dist & IC | 0.82 | 0.8078 |
| **WEST** | Weighted Edge | - | 0.8350 |

# Outline

- Word Semantic Similarity and WordNet
- Concept Specification vs. Categorization
- Weighted-Edge Based Similarity Measure
- Experimental Studies
- WEST
- Conclusion and Future Studies

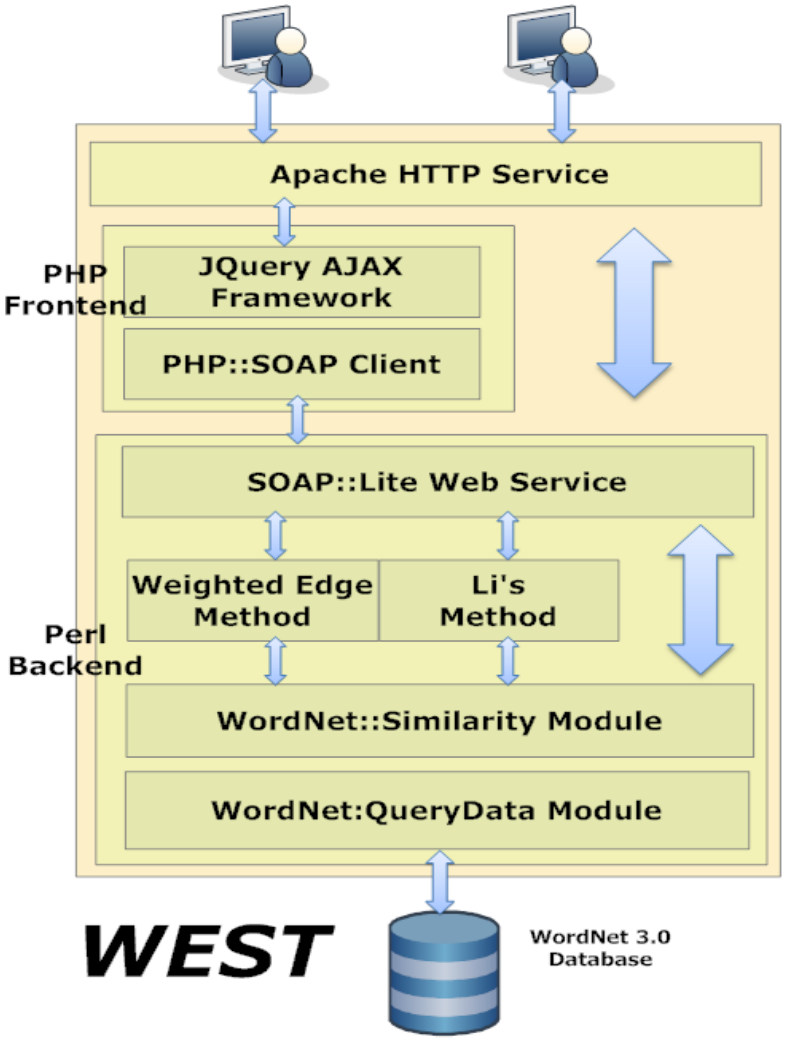
# WEST Architecture

**LCA Selection**: For each sense-pair of a word-pair, we use PathFinder from WordNet::Similarity module to retrieve its LCA

**SpecLev Retrieval**: retrieve the SpecLevs of the three target senses.

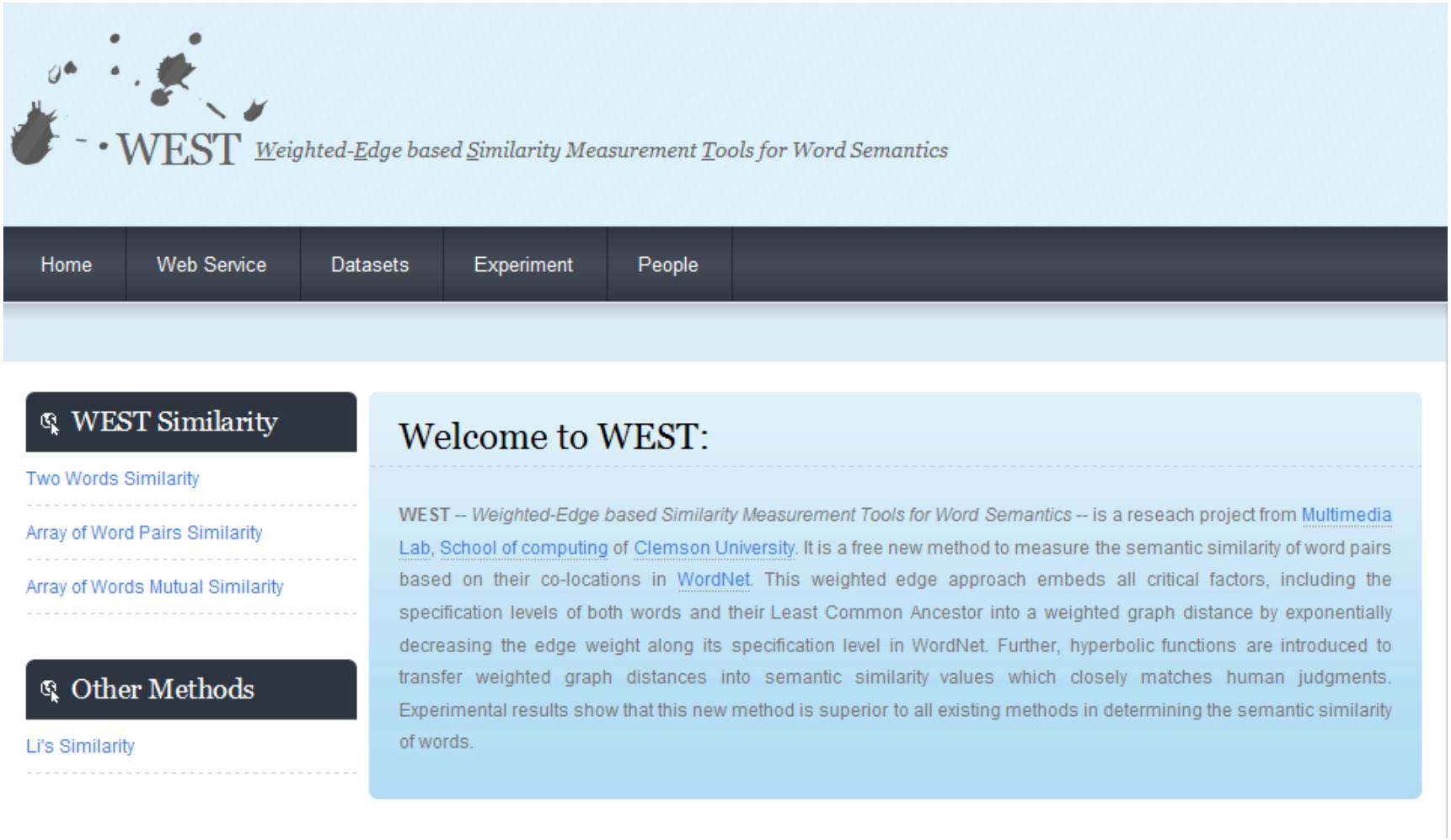
**Weighted Edge Distance**: We optimize the calculation by pre-calculating the Weighted Edge Distance from all SpecLev to its root (SpecLev 0) and store them into a 2-dimentional array for every Weighted Decreasing Rate.

**Web Service:** SOAP::Lite Module is used to wrap the Weighted Edge interface into SOAP web service.

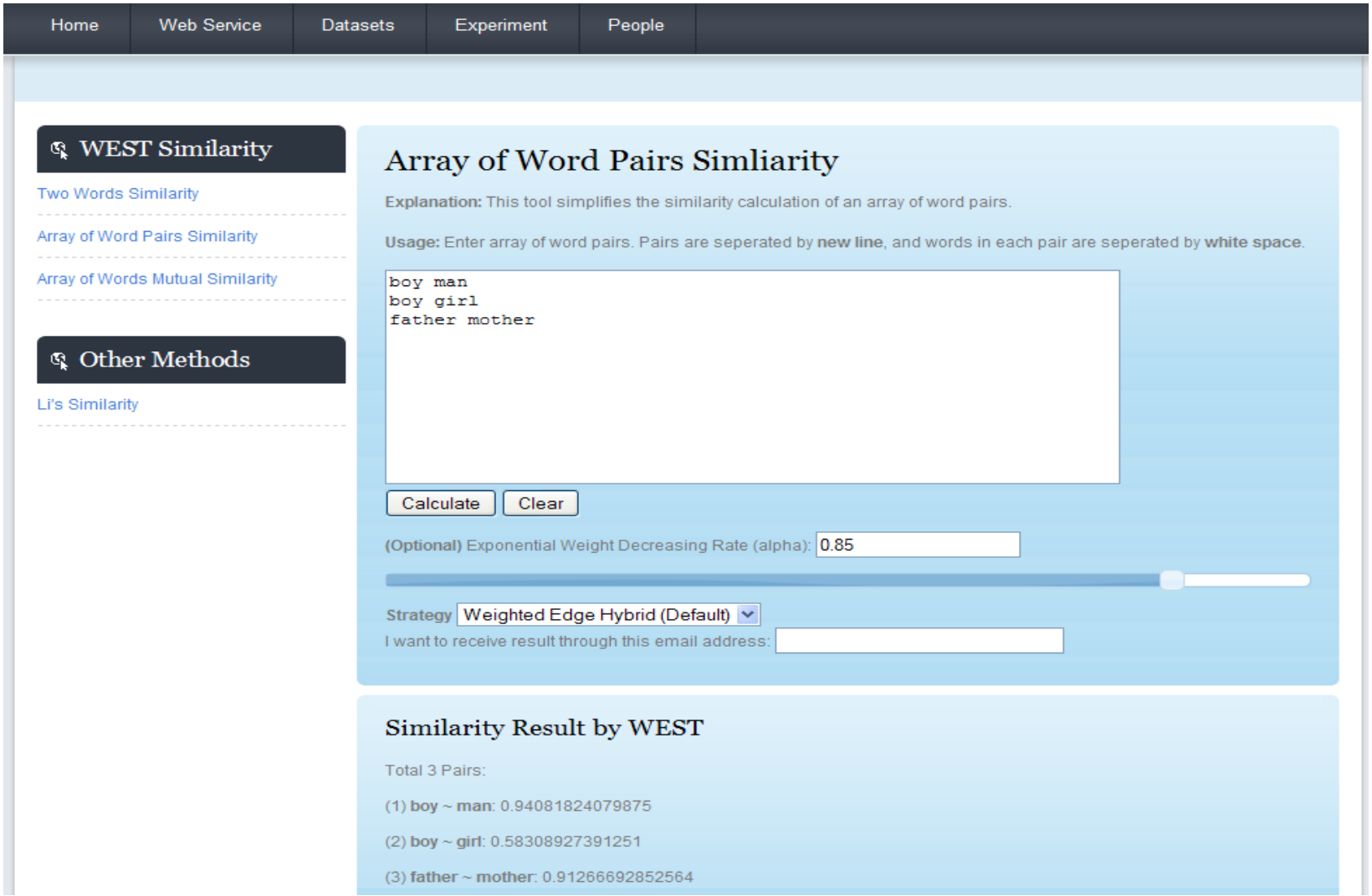


# WEST Website

**http://bioinformatics.clemson.edu/WEST/**



# WEST Website (2)



# Outline

- Word Semantic Similarity and WordNet
- Concept Specification vs. Categorization
- Weighted-Edge Based Similarity Measure
- Experimental Studies
- WEST
- Conclusion and Future Studies

# Conclusion

- *Humans are more sensitive to the semantic difference caused by categorization than specification.*
- *A novel weighted edge approach embedding the specification levels of both words and their Least Common Ancestor (LCA) into a weighted graph distance by exponentially decreasing the weight along its specification level.*
- *Experimental studies show that the similarity value obtained by this new method closely matches human perception.*
- *Based on this new method, online tools for word similarity measure are implemented as Web services.*

# Future Studies

- Measure the semantic similarity of sentences and that of text, based on WEST.
- Adapt the WEST method to measure the semantic similarity of terms in other ontologies.
- Combination of WEST and other approaches.

# Thank you!