The Geographical Analog Engine: Weighted Euclidean and Semantic Similarity Measures for U.S. Cities

T. Banchuen

GeoVISTA Center, Department of Geography
The Pennsylvania State University
302 Walker Building
University Park, PA 16802, U.S.A.
Telephone: (1) 787 360 1672
Fax: (1) 814 238 1656
Email: txb213@psu.edu

1. Introduction

Geographical analogs, which are defined here as analogous places, have been widely applied in various fields. Many scientists have used past events to forecast climate change impacts (Glantz 1988, Meyer et al. 1998, Knight et al. 2004). Analog places can serve as controls in quasi-experimental methods (Cook and Campbell 1979, Isserman and Rephann 1995, Farrigan and Glasmeier 2002). Moreover, analogs enable scientists and policy makers to communicate and generate ideas, insights, and hypotheses (Hesse 1966, Swearingen 1987, Duit et al. 2001), for example to anticipate what a given place might be like with a significantly warmer and wetter climate

Previous work has employed only numerical similarity measures, such as Euclidean and other distance measures. Valuable information does not always come in numbers and tables. It is frequently available in formats not directly amenable to direct numerical treatments, if it is available at all. Textual documents, audio/video documentaries, photographs all fit in this type of information. Such information in recent years is being generated at an accelerated rate as online blogs and Web portals (e.g., Wikipedia, YouTube, and Flickr) allow anyone with access to the Internet to contribute. We need a way to help us automatically process this wealth of information and gain new ideas, perspectives, and insights. This paper proposes a novel approach that measures similarity of places based on numerical as well as non-numerical aspects.

2. Euclidean Similarity

Following Pal (2004), the Euclidean distance is computed as follows. Let $P = (p_1, p_2, p_3, \ldots, p_m)$ denote a collection of *m* places. u_{ji} is the *i*th characteristic of the *j*th place. The *j*th place can be represented as a *n* dimensional vector $p_j = (u_{j1}, u_{j2}, u_{j3}, \ldots, u_{jn})$. The salience of a characteristic with respect to the comparison of places can be incorporated into the measure of similarity by applying a weight w_i $(w_i \in [0,1])$ to each u_{ji}. Each characteristic is normalized by its range. Using the notations above, one can compute a weighted Euclidean distance $d_{ab}^{(w)}$ between Place A p_a and Place B p_b in *P* as:

$$d_{ab}^{(w)} = d^{(w)}(p_a, p_b) = \left[\sum_{i=1}^{n} w_i^2 (u_{ai} - u_{bi})^2 \right]^{\frac{1}{2}} \quad (1)$$

The weighted Euclidean similarity score $s_{ab}^{ed(w)}$ is then

$$s_{ab}^{ed(w)} = 1 - \frac{d_{ab}^{(w)}}{\left(\sum_{i=1}^{n} w_i^2 \right)^{\frac{1}{2}}} \quad (2)$$

3. Semantic Similarity

This work experiments with two semantic similarity scores. One is the algorithm developed by Mitra and Wiederhold (2002), and the other is an enhanced version of the first.

3.1 Mitra and Wiederhold's Algorithm

The algorithm takes a textual description $desc_a$ of p_a and a textual description $desc_b$ of p_b as inputs. Given that the lengths of $desc_a$ and $desc_b$ are *q* and *p* words, respectively, the algorithm constructs a *q*-by-*p* similarity matrix of all possible pairs of words in the descriptions. s_{kl}^{sem} denotes a matrix entry for the *k*th word w_k^{desp} in $desc_a$ and the *l*th word w_l^{desp} in $desc_b$. If $p \geq q$, the similarity score s_{ab}^{sem} of p_a and p_b is computed as

$$s_{ab}^{sem} = \frac{\sum_{k=1}^{q} \max(s_{kl}^{sem}) \; for \; \forall l}{q} \quad (3)$$

else

$$s_{ab}^{sem} = \frac{\sum_{l=1}^{p} \max(s_{kl}^{sem}) \; for \; \forall k}{p} \quad (4)$$

To compute the semantic similarity score s_{kl}^{sem} between w_{ki} and w_{lj}, the algorithm looks up the definition def_k of w_k and the definition def_l of w_l using a lexicon database, WordNet (Pederson et al. 2004). It then creates an *m*-by-*n* similarity matrix between words in the definitions, where *m* is the number of words in def_k and n in def_l. Assuming w_{ki} is the *i*th word in def_k and w_{lj} the *j*th word in def_l, s_{ij}^{sem} is the semantic similarity score between w_{ki} and w_{lj} and is computed in the same manner as s_{kl}^{sem}, and therefore the algorithm is recursive. If $m \geq n$, the recursive calculation can be written as

$$s_{kl}^{sem} = s^{sem}(w_k, w_l, d) = \frac{\sum_{j=1}^{n} \max(s^{sem}(w_{ki}, w_{lj}, d-1)) \; for \; \forall i}{n} \quad (5)$$

else

$$s_{kl}^{sem} = s^{sem}(w_k, w_l, d) = \frac{\sum_{i=1}^{m} \max(s^{sem}(w_{ki}, w_{lj}, d-1)) \; for \; \forall j}{m} \quad (6)$$

while d is the recursion depth.

When $d = 0$, the computation becomes simple string matching. The algorithm first stem w_{ki} and w_{lj} to their roots with the Porter's stemmer (Porter 1980). $s_{ij}^{sem} = 1$ if w_{ki} is identical to w_{lj}, and $s_{ij}^{sem} = 0$ otherwise.

3.2 Enhanced Mitra and Wiederhold's Algorithm

It is well known in the field of information storage and retrieval that many words in documents do not help distinguish one document from another since they appear frequently in most documents (Korfhage 1997). These are called *stop words*. The enhanced algorithm purges stop words from an input text. It uses the list of stop words provided with WordNet Version 2.1 (Pederson et al. 2004).

Furthermore, humans learn and recognize concepts by associating new observations with known concepts or by classifying them into categories (Lakoff 1990). The enhanced algorithm imitates such a cognitive process by computing similarity scores between each word in an input text and those in a predefined list of categories according to the original algorithm. Any word with a similar score less than a set limit is removed from the input text, leaving behind only words comparable to predefined categories. After these two removal steps, the computation continues like the original algorithm.

4. Comparison of Six University Cities

Six U.S. university cities were selected based on the author and his advisors' judgment that their economies are highly dependent on a university or universities located there as shown in table 1. Fig. 1 displays the locations of the cities, which scatter from the West Coast to the East Coast. The comparison involves two kinds of data: statistical and textual. The statistical data considered in this comparison are taken from the *County and City Data Book* (U.S. Census Bureau 2000). Fifteen variables were chosen to represent land area, demography, economy, and climate. These topics are common for general interests of place comparison. Various comparison objectives require different sets of variables. An analyst can add or take away any variables at will. It will not affect the Euclidean similarity calculation.

City	University
Boston, MA	Massachusetts Institute of Technology
	Harvard
Boulder, CO	The University of Colorado at Boulder
Madison, WI	The University of Wisconsin
Palo Alto, CA	Stanford University
State College, PA	The Pennsylvania State University
Tuscaloosa, AL	The University of Alabama

Table 1. The six university cities and their universities.

Figure 1. Locations of the university cities

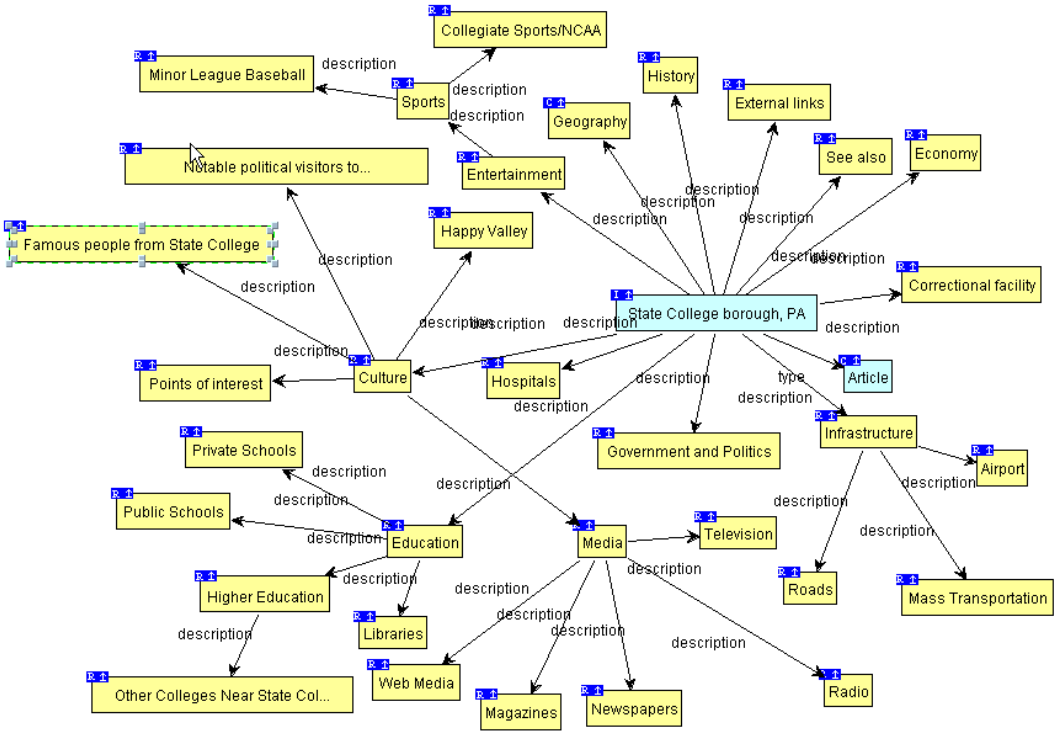

Figure 2. Concept map of the Wikipedia article of State College, PA.

The textual descriptions come from the Wikipedia articles for each city. For example, the Wikipedia article of State College, PA is available at http://en.wikipedia.org/wiki/State_College%2C_PA. Each article is parsed and converted into a Web Ontology Language (OWL) document. The parsing is guided by the table of content (TOC) of an article. A TOC entry becomes a resource in an ontology, and a subsection becomes the description of its supersection. The text in the section corresponding to an entry becomes the description of the resource. Fig 2. illustrates the created ontology of State College, PA as a concept map.

4.1 Comparison with Mitra and Wiederhold's Algorithm

The comparison employs the equal weight Euclidean similarity ($w_i = 1$) to measure similarity among cities in the statistical space. For the semantic space, the comparison always employs the Mitra and Wiederhold's algorithm to compute the semantic similarity score between two texts. The semantic similarity score between two articles is computed by dividing the sum of similarity scores of all sections in one Wikipedia article and their corresponding sections in another by the number of sections of the former. The corresponding section of a section is determined by the highest semantic similarity score between the header of that section and all headers of another article of the same level. Overall semantic scores between articles are not symmetric since one article is likely to have more sections than the other. Both the sum of section scores and the dividing section number can be different.

Fig. 3 shows the plots of Euclidean scores versus semantic scores. The title of each plot specifies the city the other cities are compared to. The Euclidean scores indicate that these cities are all quite similar when we consider all the selected statistical variables. It is probable that if we consider only subgroups of these variables, the cities would not be so much a like. However, such analysis has not yet been conducted.

The overall semantic scores, on the other hand, show that these cities are dissimilar; none of the scores is above 0.4. However, a closer look at scores between sections (not shown in the plots), reveals that some sections are almost identical ($s_{ab}^{sem} \geq 0.8$). Nonetheless, a literate person will find that the algorithm can identify a false match. For instance, consider the demographics sections of
(1) Boston, MA: http://en.wikipedia.org/wiki/Boston%2C_MA#Demographics; and
(2) Boulder, CO: http://en.wikipedia.org/wiki/Boulder%2C_CO#Demographics.

The writers of these sections follow almost identical templates. Even though the statistical values are drastically different, the algorithm will not be able to detect them since most of the words are identical. Mitra and Wiederhold's algorithm consider a number as a string. It does not know whether two numbers are about the same or not.

A review of lesser, but significant, matches suggests a further problem. The algorithm determines that the history sections of
(1) State College, PA: http://en.wikipedia.org/wiki/State_College%2C_PA#History; and
(2) Madison, WI: http://en.wikipedia.org/wiki/Madison%2C_WI#History.
have about 41 percent synonymous words. A literate person will find that they are not at all similar. This can be attributed to stop words. When stop words are removed from the text, only 14 percent synonymous words remain.

4.2 Comparison with Enhanced Mitra and Wiederhold's Algorithm

So as to patch up the flaws mentioned in the previous sections, Mitra and Wiederhold's algorithm is enhanced by purging out stop words and using predefined categories. Section 3.2 explains how these are done. As a test of concept, this work uses the census categories of economic activities (Nelson 1955) to compare only the economy sections. Similar sets of categories can be found for other sections, but are left for future work. The calculation of Euclidean similarity stays the same as the previous section and will not be discussed here.

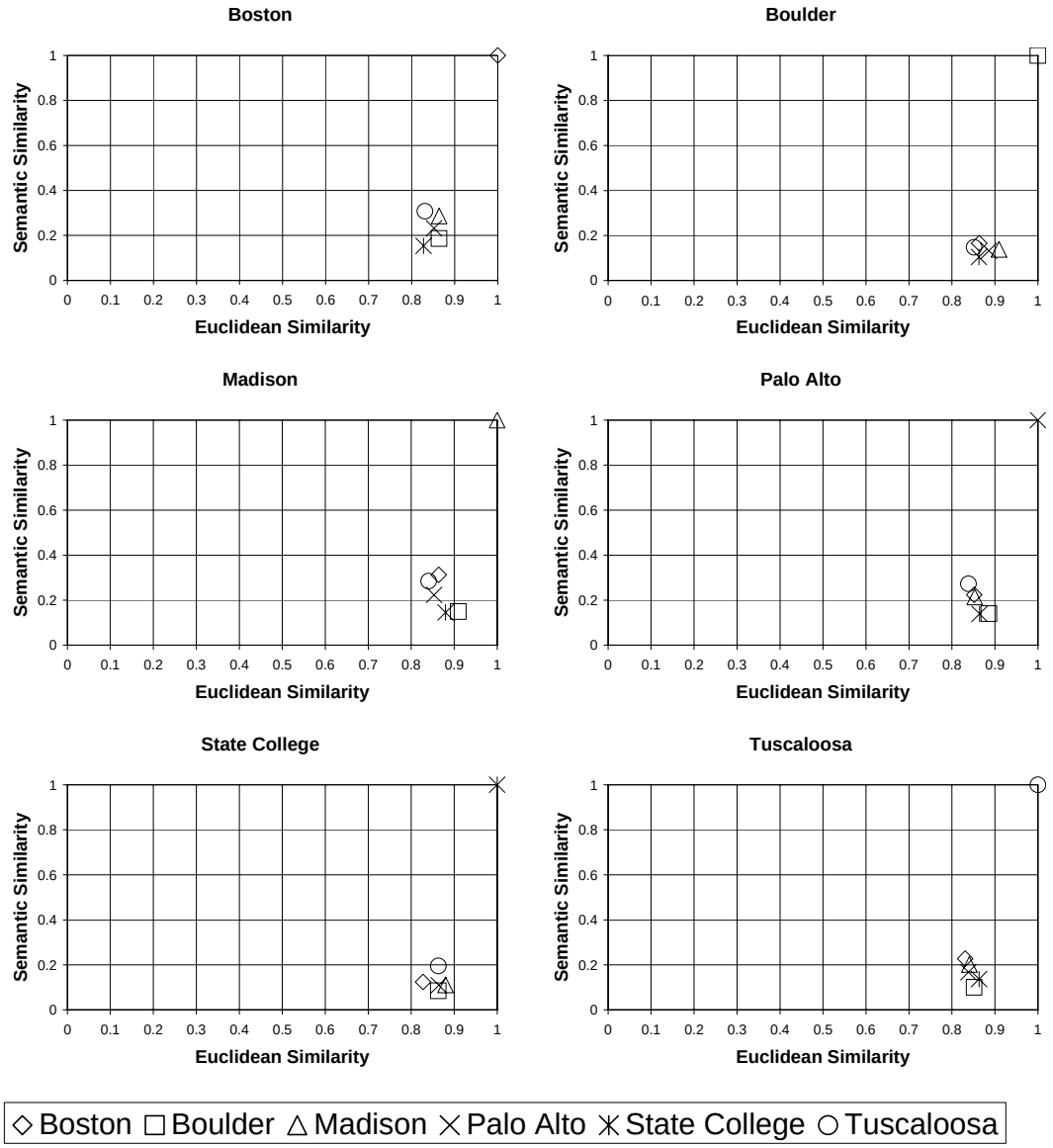

Figure 3. Plots of similarity scores computed with the original algorithm.

The enhancements produce significantly better results. The plots of Euclidean and semantic similarity scores are as shown in fig. 4. Except for Boulder, CO and Palo Alto,

CA; the enhanced algorithm produces a significantly wider range of scores, showing a better discriminating power. Boulder, CO does not match any other cities according to the economy sections. A literate person reviewing the best matches will find that they are quite good. Take for example
(1) Boston, MA: http://en.wikipedia.org/wiki/Boston%2C_MA#Economy; and
(2) Madison, WI: http://en.wikipedia.org/wiki/Madison%2C_WI#Economy.

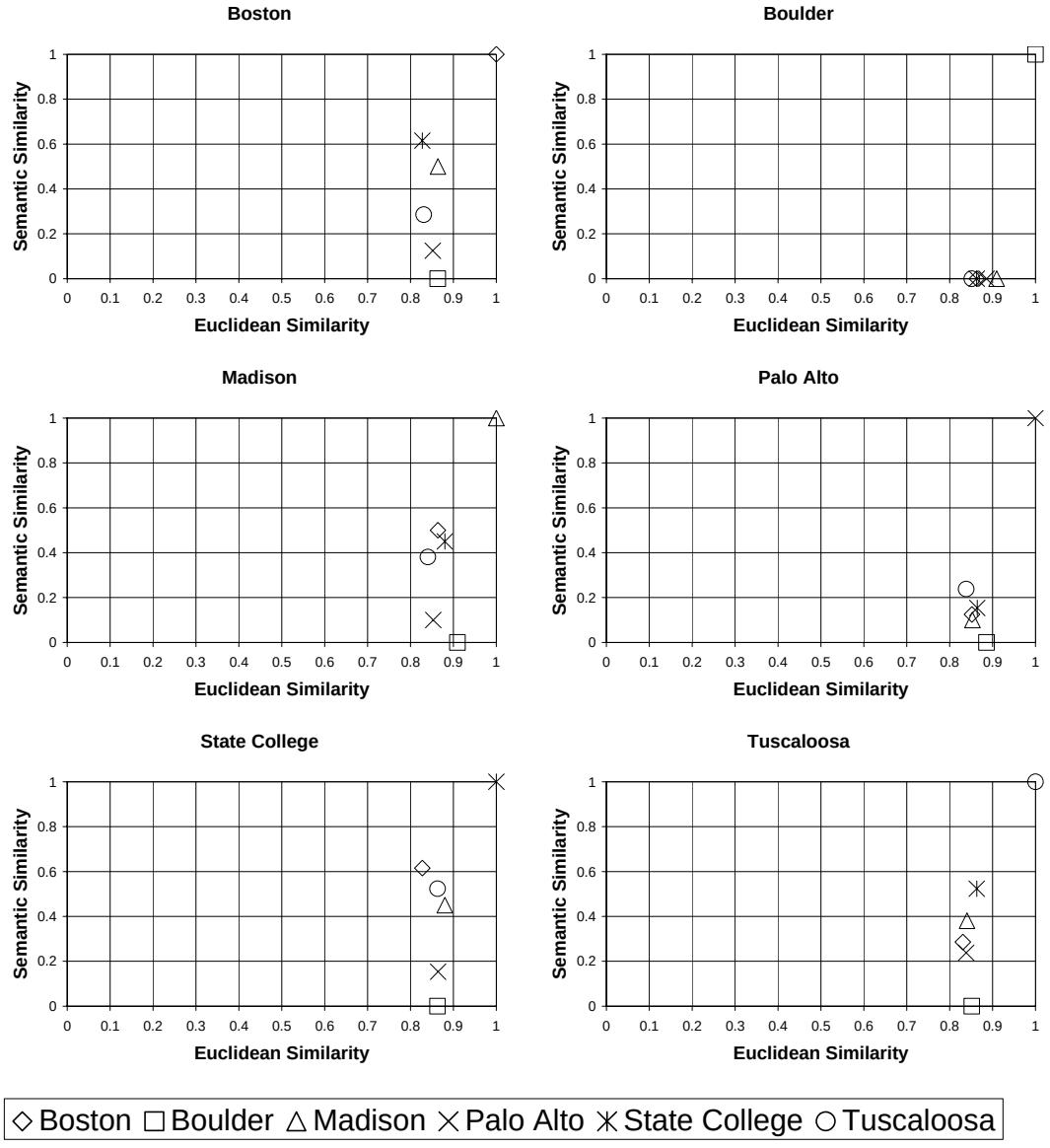

Figure 4. Plots of similarity scores computed with the enhanced algorithm.

5. Conclusions

Previous work on place analogs fails to account for non-numerical information. It has been demonstrated here that non-numerical information can be integrated automatically into identification of analogs. The novel methodology proposed provides a basis for

many more combined numerical and semantic measures. Such measures allow more comprehensive evaluation of analogs and take advantage of today's fast growing online resources.

Brute-force synonym counting as done in the first trial without predefined categories does not provide a good similarity measure. An algorithm should have some knowledge of the domain it is computing about in order to yield meaningful similarity scores. Future research that attempt to incorporate several other existing techniques from information science, such as taking into account word proximity and part-of-speech should prove fruitful. Two synonyms produce different meanings if they appear in one sentence in one article and in separate sentences in another. One word can have various senses; detecting what sense it has in a sentence will surely increase the accuracy of a semantic similarity algorithm. Note that correctly recognizing meaning of numbers in semantic similarity calculation may not be so important since the methodology already analyzes statistical data.

Semantic similarity calculation is very computationally expensive. It takes about 3 to 6 hours to compute this small example on a Pentium III machine, compared to a few seconds for Euclidean similarity on the same machine. Hence, for a larger set of cities, one can try to speed up the calculation by only computing semantic similarity for the top ten or so places according the Euclidean similarity.

6. Acknowledgements

Special thanks to the National Science Foundation and the GeoVISTA Center, the Geography Department at the Pennsylvania State University for providing financial support of this research. The author would also like to thank Dr. Mark N. Gahegan and Dr. C. Gregory Knight for their valuable advice and mentoring.

7. References

U.S. Census Bureau, 2000, *County and City Data Book*, U.S. Census Bureau, Administrative and Customer Services Division, Statistical Compendia Branch.

Cook, Thomas D. and Campbell D T, 1979, *Quasi-Experimentation: Design & Analysis Issues for Field Settings*, Houghton Mifflin Company, Boston, MA.

Duit R., Roth W-M, Komorek M and Wilbers J, 2001, Fostering conceptual change by analogies between Scylla and Charybdis, *Learning and Instruction*, 11 (4-5):283-303.

Farrigan T L and Glasmeier A K, 2002, The economic impacts of the prison development boom on persistently poor rural places, The Earth and Mineral Sciences Environmental Institute, Pennsylvania State University, State College, PA.

Glantz M H (ed), 1988, *Societal Responses to Regional Climatic Change: Forecasting by Analogy*, Westview Press, Boulder, CO.

Hesse M B, 1966, *Models and Analogies in Science*, University of Notre Dame Press, Notre Dame, IN.

Isserman A and Rephann T, 1995, The economic effects of the Appalachian Regional Commision, *Journal of the American Planning Association*, 61 (3):345-364.

Knight C G, Raev I and Staneva M P, 2004, *Drought in Bulgaria: A Contemporary Analog for Climate Change*. Ashgate Publishing Company, Burlington, VT.

Korfhage R R, 1997, *Information Storage and Retrieval*, John Wiley and Sons, Inc, New York, NY.

Lakoff G, 1990, *Women, Fire, and Dangerous Things*, University of Chicago Press, Chicago, IL.

Meyer W B, Butzer K W, Downing E T, Turner II B L, Wenzel G W and Wescoat J, 1998, Reasoning by analogy. In Rayner S and Malone E (eds), *The Tools for Policy Analysis*, Battelle Press, Columbus, OH.

Mitra P and Wiederhold G, 2002, Resolving terminological heterogeneity in ontologies, In: *Proceedings of ECAI-02 Workshop*, CEUR-WS 64.

Nelson H J, 1955, A service classification of american cities, *Economic Geography*, 31 (3):189-210.
Pal S K, 2004, *Foundations of Soft Case-Based Reasoning*, Wiley-Interscience, Hoboken, NJ.
Pederson T, Patwardhan S and Mechelizzi J, 2004, Wordnet::Similarity - measuring the relatedness of concepts. In: *Proceedings of the National Conference on Artificial Intelligence*, July 25-29, San Jose, CA.
Porter M,1980, An algorithm for suffix stripping, *Program*, 14 (3):130-137.
Swearingen W D, 1987, *Morrocan Mirages*, Princeton University Press, Princeton, NJ.