Discrete Mathematics Seminar

Time: Friday, 19 February 2010, 12:30-1:30 PM
Room: 238 Derrick Hall
Title: Graph theory applications to Google search
Speaker: Dr. Ralucca Gera, Department of Applied Mathematics, Naval Postgraduate School

ABSTRACT:

One of the chief concerns of linguists is the ambiguity of natural language. At the lexical level, this manifests in the existence of the multiplicity of senses that a word may have. A natural representation for the outcome of this procedure is a graph, where V is the vocabulary (the set of distinct words in the text) and vertices are adjacent in G if the words they represent co-occur in a relevant pattern in the text. Ideally, the words in the same semantic field thus give rise to a component of the graph. However, when words that have multiple senses are part of the graph, this is not the case.

In response, Dorow et al. provide a technique that transforms the graph into a new graph, for which generally each individual component contains only one meaning of the polysemous words. This new technique, called link graph (similar to the line graph), will help to automate discovery of ambiguous words. This transformation is identical to the triangular line graph, a special case of the H-line graph introduced by Chartrand et al: the triangular line graph of G, denoted by T(G), is the graph with vertex set E(G), with two distinct vertices v_e and v_f adjacent in T(G) if and only if there exists a subgraph $H \cong K_3$ of G with $e, f \in E(H)$. The properties of the T transformation have been studied by Jarrett for K_n and Dorrough for arbitrary G, with emphasis on stabilization of iterations of $T(K_n)$ and T(G), respectively. This presentation summarizes the main known results, and examines how the structural properties of triangular line graphs can aid predictions of the curvature metric on the triangular line graph, thereby helping to identify polysemous words.

We applied this method to the graph obtained from the Gigaword corpus, and to our surprised we obtained one component with over 50% of the words not distinguishing the meanings of the words. And so, we also propose to use one other measure on link graphs, and our evaluation of this method resulted in removal of 15% of the edges in the original graph from the Gigaword corpus. Application of the link graph transformation to this new trimmed graph resulted in 900 components, with the largest containing only 5% of the words in the data set, a major division of what occurred without this procedure.

Co-authors: Pranav Anand (Linguistics Department, University of California Santa Cruz), Henry Escuadro (Department of Mathematics, Juniata College), and Craig Martell (Computer Science Department, Naval Postgraduate School).