

Blake, C. (2007) In Support of e-science - Shifting from Information Retrieval to Information Synthesis, Presentation at the 2007 Microsoft eScience Workshop at RENCI, Chapel Hill, NC.

Abstract: Cyberinfrastructure provides students, scientists and policy makers with an unprecedented quantity of information, for example in biomedicine, PubMed adds 12000 new citations each week and the top chemistry journals publish more than a hundred thousand articles in a single year. Despite advances in information access, the quantity of information far exceeds human cognitive processing capacity. Consider a breast cancer scientist who must sift through the 12600 articles published during the 28 months required to conduct a systematic review, a process used to resolve conflicting evidence. In addition to quantity, evidence related to the complex research questions posed by scientists transcend traditional disciplinary boundaries and thus require a multi-, inter-, or trans- disciplinary approach.

I will describe how recent advances in natural language processing, specifically in recognizing textual entailment and in generating multi-document summaries can enable new kinds of e-science. The next generation of information tools will recognize contradictions and redundancies that are inevitable in the information intensive environment in which a scientist operates. Using existing systems that account for complex interdependencies between scientific articles as examples, I will show how these systems embody the shift from the information retrieval to information synthesis. I will conclude with preliminary results from Claim Jumper, a system that captures the spirit of gold-miners searching for nuggets of knowledge in a new frontier, and reflects a scientist's transition through traditional disciplinary boundaries. Given a topic, query and set of articles, Claim Jumper generates a fluent well-organized summary from published literature that accounts for redundancy.