MetaSearch: Data Fusion and Distributed Retrieval

Javed Aslam
Jamie Callan
R. Manmatha
Mark Sanderson
Ellen Voorhees

Draft: November 4, 2002

1 Motivation

MetaSearch is the process of retrieving and combining information from multiple sources, and it is typically studied in one of two forms: (1) data fusion, the combination of information from multiple sources that index an effectively common data set and (2) collection fusion or distributed retrieval, the combination of information from multiple sources that index effectively disjoint data sets.

As more and more retrieval mechanisms become available over common data sets (e.g., the World Wide Web) and specialized data sets (e.g., medical and law libraries), the process of identifying likely sources of relevant information, retrieving information from those sources, and effectively combining the information thus gathered will only grow in importance. A future wherein ubiquitous mobile wireless devices exist, capable of forming ad hoc peer-to-peer networks and submitting and fielding requests for information, gives rise to a new host of challenges and potential rewards.

2 Distributed Retrieval

The issues typically addressed by a distributed retrieval system include: (1) resource description, i.e., describing the data set contents, services available, query syntax, etc.; (2) resource ranking, i.e., ranking multiple resources by their likelihood of containing relevant information; (3) resource selection, i.e., selecting the “best” subset of resources from a ranked list; (4) searching, i.e., retrieving information from the selected set of resources; and (5) merging, i.e., combining the retrieved information from multiple resources.

Many of the techniques developed to address these issues are ad hoc in nature, though language modeling techniques have been successfully employed to address resource description, ranking and selection.
2.1 Three to five year challenges

(1) Can a standard resource descriptor be devised such that if a resource published its descriptor (e.g., on the web), it could participate in a generic distributed retrieval system? What data must be present in such a descriptor? A language model of the underlying data set? A semantic description of the content? A model of the required query syntax which permits interoperability?

(2) The performance of language modeling techniques is, at present, on par with that of *ad hoc* techniques. Can a theoretically grounded model of distributed IR be developed which consistently outperforms *ad hoc* techniques?

(3) The performance of distributed IR techniques is approaching that of a “single database,” at least within research environments. Can this be achieved in practice? Furthermore, through the judicious use of resource selection, distributed IR should, in theory, outperform a “single database.” Can this be achieved?

3 Data Fusion

The task of a metasearch data fusion system is to combine the results of multiple search engines run over an effectively common data set in response to a given query. Classic techniques for this problem most often assume that relevance scores are available from the underlying search engines, and these techniques typically address the following issues: (1) *relevance score normalization*, i.e., mapping the relevance scores given by multiple search engines to a common (and comparable) space and (2) *normalized score combination*, i.e., obtaining a final score for each document from the normalized scores, from which a final ranking may be obtained. Common techniques for relevance score normalization include (1) linear mapping to a fixed range (e.g., \([0, 1]\)) and (2) score distribution normalization (e.g., shift and scale to achieve a common mean and variance). Normalized scores are typically combined using a (weighted) linear average.

Other data fusion techniques that have been developed include (1) modeling the problem as a multi-candidate election and employing rank-aggregation algorithms from Social Choice Theory (e.g., the Borda Count and Condorcet methods) and (2) various supervised learning techniques (e.g., boosting, cranking, naive Bayes, etc.).

3.1 Three to five year challenges

(1) The performance of search engines vary from query to query. The goal of metasearch is often to outperform the *(a priori* unknown) best underlying search engine on a per query basis, and this can typically be achieved when combining systems of similar performance. However, this goal is often unachieved when combining search engines of widely varying levels of performance. Can a metasearch technique be developed which consistently outperforms the best underlying search engine? Or can a technique be developed which is capable of distinguishing the “good” underlying systems from “bad” on a per query basis?
(2) Techniques for data fusion typically assume (and often implicitly require) that the underlying search engines index an effectively common data set; techniques for distributed IR typically assume that the underlying search engines index effectively disjoint data sets. Can techniques be developed which effectively combine the results of underlying search engines that index data sets of varying and unknown overlap? (Search engines on the web fall within this category.) Can the metasearch problem be modeled in a unified way such that data fusion and collection fusion are merely two extremes of a single underlying problem?

4 Grand Challenges

4.1 Massively distributed multi-lingual retrieval

On the World Wide Web alone, digital information is being accumulated at a phenomenal rate and in a vast number of languages. By some estimates, the number of web pages written in Mandarin Chinese will overtake those written in English in the not too distant future. A grand challenge of IR will be to develop massively distributed, multi-lingual retrieval systems. Such a system would take as input an information need, encoded in any language, and return relevant results, encoded in any language. The design of such a system would probably borrow from techniques in distributed retrieval, data fusion, and cross-lingual IR, but it is unlikely that simply combining known techniques from these fields would yield such a system. What would the architecture of such a system be? What data sets could be used to develop and test such a system?

4.2 Peer-to-peer ad hoc IR

Wireless devices are being rapidly developed which can spontaneously form ad hoc peer-to-peer networks. A world in which such devices are ubiquitous gives rise to a number of challenges and potential rewards for IR. For example, one might fancifully imagine flying to an unknown city for a business trip, renting a car, opening one’s wireless PDA, and connecting to an ad hoc peer-to-peer network in order to submit the following information requests: (1) directions to the nearest gas station from the current location, (2) recommendations for an Italian restaurant with five miles of a specific hotel, and (3) a current weather forecast for the region. How would such information needs be requested? (Free text, structured queries, etc.) To whom would such requests be sent? (Some requests, such as regional weather, could be forwarded to central servers such as Yahoo!; others, such as restaurant recommendations or gas station directions, could be answered by those establishments or by other users on the network with such knowledge.) How would the results obtained be combined and presented to the user? (Perhaps an “e-bay”-like trust model could be employed as a filter to obtain “trusted” content and avoid SPAM.) What would the architecture of such a system be, and what data sets could be used to develop and test such a system?