To what problem is distributed information retrieval the solution?

Paul Thomas

CSIRO

GPO Box 664, Canberra 2601, Australia

paul.thomas@csiro.au

Abstract

Distributed information retrieval, where a single broker coordinates retrieval from many independent search services, has been extensively studied—but typically without any particular application and sometimes even without any explicit motivation. There have been a handful of arguments given for distributed IR—coverage, effectiveness, and ease of use for example—but these are not borne out by experience.

I ask: are there in fact uses for distributed IR? There are, but generally for organisational, not technical, reasons, and they have not been well-studied.

Distributed IR

Distributed information retrieval, or DIR, combines several independent search engines into a single interface. A *broker* communicates on the user's behalf with one or more engines, coordinating parallel searches and presenting a single set of results (see Figure 1). In the *uncooperative* model (Callan, 2000), brokers and servers communicate only via a standard query/results interface. This model has been extremely popular in the research literature if only modestly popular in "real world" application.

[Figure 1 about here.]

Brokers typically incorporate several tasks: discovering available sources; characterising these by language, topic, or other attributes; routing a user's query; translating between the broker's query language and that of each source; parsing results from each engine; merging results into a single list; and finally presenting that list. Recent reviews (Meng et al., 2002; Shokouhi and Si, 2011) are comprehensive in their coverage of techniques for each step. They do not, however, devote much space to an important question:

When is distributed IR actually useful?

For example, despite claims to the contrary it is not clear that distributed IR is needed for coverage. Nor is it likely to improve results over those from a conventional system.

Markov (2011, in oral presentation) has referred to "academic" distributed IR, where techniques are advanced and evaluated in the absence of any particular application. This has been very useful for developing algorithms, but should be put in the context of applications. This would not only shed light on the effectiveness of existing systems, but also on how we should test distributed IR systems; how much scope there is for improvement; and where to direct research effort.

Do we really need an application here and now? It is possible to proceed anyway, and to trust that research will eventually pay off. Understanding a good application does not preclude working on other cases—or no case at all—but this should be a conscious choice. There is no point solving non-problems, so at very least we need to be sure we are not wasting our time. A well-developed use case may also suggest new research problems.

A note on terminology The phrase "distributed retrieval" is also used to describe a similar idea, distributing a (conceptually) single search engine across many nodes—geographic distribution, rather than distribution of control. This problem is clearly related, but there are several key differences. There is, in theory, complete information about each node at the broker; there is complete cooperation with the broker; and neither discovery, characterisation, query translation, nor result parsing is needed. Geographic distribution is important for global-scale search engines, but it is not clear that this is an instance of the conventional DIR model.

Applications outside distributed IR Techniques from DIR, especially selection and merging, have been applied elsewhere. In aggregated search (Lalmas and Murdock, 2008), the task is to integrate information of several types—for example, maps, news, and the web—and algorithms from distributed IR have been used to choose amongst sources (see e.g. Arguello et al., 2009). Blog and "expert" search engines have also drawn on algorithms from distributed IR (see e.g. Elsas et al., 2008). Again individual algorithms are useful, but in principle all data is locally available. These are not instances of the standard uncooperative model.

Claims for distributed IR

Several claims have been made for uncooperative DIR, including increased coverage and improved results. I assess these claims below, but contend that they are not borne out by real examples.

Coverage in web search

Early research was motivated largely by concerns of coverage for web search: since no single search engine could cover such a large corpus, it was argued, aggregating several engines must help.

MetaCrawler (Selberg and Etzioni, 1995) illustrates the argument. It combined six search engines to increase coverage of web search, since "no single search service [was] sufficient" and at the time it seemed unlikely that a single service would ever provide enough coverage. Selberg and Etzioni calculated that the best contemporary search engine could return only 45% of useful results, and concluded that distributed IR significantly improved web search.

Other systems have included ProFusion (Gauch et al., 1996), SavvySearch (Dreilinger and Howe, 1997; Howe and Dreilinger, 1997), Inquirus 2 (Glover et al., 1999), and Anvish (Shu and Kak, 1999). All were motivated by the apparent impossibility of a single engine covering the web, and all performed well in evaluations. As of early 2012 the model is still alive and well, with systems including Dogpile, Ixquick, and Mamma.

This argument is supported by a number of studies on search engine coverage (Bharat and Broder, 1998; Lawrence and Giles, 1998; Gordon and Pathak, 1999; Lawrence and Giles, 1999; Bar-Yossef et al., 2000; Henzinger et al., 2000; Gulli and Signorini, 2005; Bar-Yossef and Gurevich, 2006). Although difficulties in sampling make individual estimates unreliable (Thomas and Hawking, 2007), the overall result is clear: no search engine indexes the full web, and different engines index different parts.

The argument for coverage is still explicitly repeated (see e.g. Spink et al., 2006; Sadeghi, 2009). But is coverage still a concern? The major search engines index tens of billions of pages, which certainly suggests good coverage of a range of topics. We can assume that crawls are organised to get important pages first: pages which are missing from a large index are not likely to be popular targets. The market dominance of Google and Bing, compared with distributed engines such as Dogpile, certainly suggests most users are not worried about coverage.

There is also little evidence that increased coverage helps web search. Besides the experiment of Gauch et al. (1996) there are few studies addressing effectiveness, and those do not distinguish the effect of coverage from other factors such as user interface design. Limited coverage may have impaired effectiveness in early web search, but by 1999 there was no correlation between performance and index size (Hawking et al., 2001). It appears that there has been no more recent study of these factors, but coverage is no longer a convincing argument for distributed web search.

Improved results

A number of studies have shown distributed retrieval can be more effective than retrieval from a central index. The improvements are modest, and highly dependent on both the documents and the algorithms used—in many cases, needing "oracle" components—but they have been seen repeatedly (see for example Hawking and Thistlewaite, 1999; Xu and Croft, 1999; Craswell et al., 2000; Powell et al., 2000; Rasolofo et al., 2001; Abbaci et al., 2002; Thomas and Shokouhi, 2009). The improvements can be explained by the cluster hypothesis and by source selection: relevant documents will tend to be at the same source, and only sources with several relevant documents will be selected.

Even these small improvements have not been seen in practical systems, however. Using data from 1999, Hawking et al. (2001) compared the effectiveness of twenty search engines on a variety of metrics. The two distributed engines in the survey—Inquirus and MetaCrawler—were no more effective than the traditional designs. More recent work by Kumar and Pavithra (2010) compared two distributed engines (MetaCrawler and Dogpile) with two conventional engines (Google and Yahoo!). They also saw no difference in precision; surprisingly, the two distributed engines performed poorly on recall.

A distributed IR system could in theory be more effective than a conventional alternative. However, the gain is very modest, and it seems highly unlikely that we will see any improvement in practice.

Ease of use

A third common argument is that a broker's unified interface can simplify users' tasks. When each collection instead presents its own search interface, a user must use several tools to find information they need. As well as remembering how to operate a separate tool for each collection, and the capabilities of each, a user must decide which collection is the most appropriate for each task.

The requirement that the user choose an engine for each query also makes certain errors possible. The first occurs when an engine is chosen, but nothing relevant is returned; a user may give up at this point, unaware there is relevant information elsewhere. The second occurs when a user chooses an engine and finds apparently relevant information; however, better information may be available elsewhere (Thomas, 2008).

This argument is illustrated by the MetaSEEk tool of Beigi et al. (1998), which offered DIR for images. MetaSEEk was motivated by the difficulties of "knowing where search engines are, what they are designed to retrieve, and how to use them". Evaluation (Benitez et al., 1998) suggested MetaSEEk users could find a target image faster than if they used an unnamed, simple broker with random "selection"; however, earlier experiments had shown no difference in explicit user feedback even against the random baseline (Beigi et al., 1998). No evaluation compared the distributed approach with a single index.

The growth of desktop search tools, and aggregated search for the web, certainly suggests appetite for a single search box. Normally, however, "desktop" just means data stored in the local filesystem, and there is no need for distributed IR in the current generation of desktop search tools. DIR could add remote sources in a unified search interface, but no work has established whether this would improve usability.

There are also usability costs associated with distributed IR. It is possible that a user may remember where they saw something, or what tool they used to find it (Teevan et al., 2004); these clues would be lost if every search looked the same. Some users also care about provenance of results and the features of individual search engines. This is a particular issue in medical search, legal discovery, patent search, or archival search for example.

There is a tradeoff. It seems likely that DIR tools will improve ease of use overall, especially if they allow control of individual sources. However, there is no firm evidence for this, and little attention has been given to case studies or to designing usable interfaces (although see Park (1999) or Wrubel and Schmidt (2007)).

Fast-changing collections

Results returned from an out-of-date index may miss recent additions, may be based on old statistics, or may include documents no longer in the collection. Distributed IR may help to overcome this problem: each collection is indexed by its own engine, in response to updates or at least fairly frequently, and by aggregating searches we allow good coverage with reduced staleness.

By 2007, the major web search engines were re-indexing changed pages in a day or two, and Google was adapting its crawling frequency for each page (Lewandowski, 2008). Techniques from distributed IR are therefore unlikely to help in general web search, but might be useful where updates happen more than daily or where crawling and indexing is particularly slow.

News search provides an example. The assumption is that news agencies will keep their own systems up to date; a distributed IR system can then produce up-to-the-minute results without re-crawling. PENG, a system for journalists (Baillie et al., 2006), and AllInOneNews, a system for general users (Liu et al., 2007), illustrate the idea.

News aggregation would seem an obvious application for distributed IR, but experiences have been mixed. Comments from researchers involved with PENG suggest that a traditional model would have been a simpler design and may have been more effective (personal communication).

Evaluations of AllInOneNews, comparing it with Mamma News (also distributed) and Google News (centralised), were more positive. Google was better at duplicate removal and diversity, but the engines were similar on average precision (Liu et al., 2007). AllInOneNews did perform better on a recency-biased form of precision, but it is possible that the differences are due to differences in ranking not coverage. (Mamma News performed poorly, but the authors note "[t]he reason why Mamma News did not do well in this regard probably has something to do with its result-merging algorithm". The same could surely be true of Google News.) Because of this confounding factor, the argument for distributed retrieval is weak; at any rate, at the time of writing, Google News appears to be crawling news sources every few minutes.

In neither case—PENG nor AllInOneNews—was the distributed system evaluated against a central system with similar coverage.

A similar argument can perhaps be made with the growth of media such as Twitter. It may be desirable to provide results from these media, but to leave problems of real-time indexing and retrieval to others. If this is for technical reasons—that is, only the media provider can do the searching—this would be an argument for a distributed model. To the best of my knowledge, there is no public information which would make (or refute) this case.

Where might distributed IR be useful?

Distributed IR is probably not useful for the problems normally cited—coverage and improved results—and there is insufficient evidence for claims about ease of use or fast-changing collections.

Simeoni (2004) argues that if there is cooperation between broker and servers, then a single central collection should be preferred to a distributed model. His argument is sound; but there are cases where cooperation cannot be assumed and distributed IR is reasonable.

These cases are forced not by technical constraints, but by organisational and business decisions which prevent full cooperation. The consequences for IR systems are not well understood, and better understanding of information needs and sources would be most helpful.

The deep web, uncrawlable sites, private collections

The "deep" ("hidden" or "invisible") web is those documents which are online but which cannot be crawled. Commonly, this is because the only access is by a search tool, without support for browsing. Many online databases work on this model.

The deep web is estimated at 500 times the size of the "surface" web, with half a million databases and a million interfaces (He et al., 2007). Although some of this is available via conventional web search, a large majority is not (He et al., 2007; Madhavan et al., 2009). If we want to search two or more such collections at once, or we want a broker to recommend collections, distributed IR is required.

The problem is not confined to the public web. Enterprise search provides similar cases: as well as its own documents, an enterprise may have access to any number of uncrawlable resources. These may include subscription services as well as internal databases or contact lists. Access to some collections is charged for every document retrieved; access to others is restricted by licence agreements. Again if we want to search more than one collection, or want a tool to suggest places to look, we are forced into a distributed model.

I am not aware of any large-scale study of enterprise resources, but anecdotal evidence is that there are often a large number, varying greatly in holdings, access methods, and interface; that these differ from enterprise to enterprise; and that searching several at once would be useful. These assumptions need to be tested with real data, which would also help understand the sources enterprise systems should (or can) cover.

Personal search—desktop search extended to include external resources—provides a further case. This is similar to enterprise search, and findings from enterprise studies may carry over. It seems likely, however, that privacy concerns will require a broker for each user, which means they must run on smaller devices. This restricts the resources available, and constrains system designs. Again, to date there are no comprehensive studies of this setting.

If enterprise and personal search are indeed useful applications, they suggest their own research problems. Resource discovery has been largely ignored in DIR research, but will be important for large-scale deployments and for personal retrieval. An automated system would also need to characterise each search interface. Query translation will be important; so will be some way to account for different search types and differing effectiveness of each server. If servers charge for access, we will need cost-aware algorithms. Finally, in the enterprise case, different users may be allowed access to different subsets of information. Most of these problems have seen little study.

Integrating several agencies

DIR may have a role where several agencies cooperate to provide information, but where for technical or organisational reasons they cannot provide a single search service. This may be because staff time is not available, because crawling is impossible, or when the agencies want to keep control over their data. Examples include government departments, which want to provide information on a common topic but which cannot centralise their records; libraries, which want to reference each other's holdings while maintaining control of their own search; and businesses flying in loose formation, for example to cooperate on a tender, which want to retain control of their information.

In this scenario, individual search engines know something about the documents they index and the users they serve. Therefore, they can offer appropriate local optimisations: translating terms, for example, or ranking results according to popularity. This specialised knowledge would be lost if a centralised engine were used.

FedLemur (Avrahami et al., 2006) is the only full-scale example of a distributed IR system starting from this motivation. It searched official statistics from 20 US government agencies via existing interfaces. Tests reported by Avrahami et al. tuned parameters of the broker; however, no comparisons have been reported with a central index. By 2011, FedStats search was provided by Google. Other whole-of-government search engines including those of the USA, Australia, and New Zealand use a web crawler and a central index.

Integrating agencies with distributed IR is an interesting possibility, but at present it is unproven outside small case studies. It would reward further investigation. At a minimum, we need to know how often these needs arise, and how much cooperation is likely. Particular applications will present their own additional questions, for example around provenance, ranking, and security.

Directions

Table 1 summarises the discussion above. Most of the claims made for distributed IR seem unlikely, or at least are not supported by good evidence. So where next?

[Table 1 about here.]

1. We should consider how much distributed IR can contribute to aggregrated search; expert search; peer-to-peer search; search in social media; and anywhere else where (a) there is a distribution of data or (b) retrieval is of a *set* of documents, not a *single* document. There may be useful contributions—for example, in characterising or selecting resources—even if these are not DIR applications.

2. Starting with something like Table 1, it would be worthwhile to characterise applications in some detail. This will suggest which applications are worth pursuing, but also what the problems look like: what would be helpful to users, how documents can be accessed and used, what the costs are, etc.

3. Current DIR techniques will most likely be a poor fit to these applications. An evaluation of existing techniques for each new application would confirm or refute this, and inform future designs.

4. Any evaluation needs test cases—interactive experiments, test collections, techniques for log analysis, etc—which represent possible applications. Existing testbeds are unlikely to do this well. A test collection which fairly represents the deep web would be a contribution; and a technique to evaluate search over private collections would enable a lot of interesting research.

5. It will be important to explicitly compare distributed IR systems with a non-distributed option, to understand what we're getting for our trouble. This has been missing in most work, so we cannot say whether distributed retrieval really helps or hinders system effectiveness.

In particular, it seems that searches over the deep web and over personal, private, or otherwise controlled collections provide good use cases for distributed IR. A good description of these situations, a characterisation of the associated retrieval problems, and a robust set of tests for search techniques would all be valuable steps toward understanding real applications and making a real difference.

Acknowledgements I thank Mark Baillie, Nick Craswell, David Hawking, Milad Shokouhi, and Stephen Wan for their useful comments.

References

Abbaci, F., Savoy, J., and Beigbeder, M. (2002). A methodology for collection selection in heterogeneous contexts. In *Proc. IEEE Conf. on Information Technology*, pages 529–535.

Arguello, J., Diaz, F., Callan, J., and Crespo, J.-F. (2009). Sources of evidence for vertical selection. In *Proc. 32nd Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 315–322.

Avrahami, T. T., Yau, L., Si, L., and Callan, J. (2006). The FedLemur project: Federated search in the real world. *J. American Society for Information Science and Technology*, 57(3):347–358.

Baillie, M., Crestani, F., and Landoni, M. (2006). PENG: Integrated search of digital news archives. In *Proc. 29th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 607–608.

Bar-Yossef, Z., Berg, A., Chien, S., Fackcharoenphol, J., and Weitz, D. (2000). Approximating aggregate queries about web pages via random walks. In *Proc. 26th Int. Conf. on Very Large Data Bases*, pages 535–544.

Bar-Yossef, Z. and Gurevich, M. (2006). Random sampling from a search engine's index. In *Proc. 15th Int. Conf. on World Wide Web*, pages 267–376.

Beigi, M., Benitez, A. B., and Chang, S.-F. (1998). MetaSEEk: A content-based meta-search engine for images. In *Proc. 1998 SPIE Conf. on Storage and Retrieval for Image and Video Databases VI*, volume 3312 of *Proceedings of SPIE*, pages 118–128.

Benitez, A. B., Beigi, M., and Chang, S.-F. (1998). Using relevance feedback in content-based image metasearch. *IEEE Computer*, 2(4):59–69.

Bharat, K. and Broder, A. (1998). A technique for measuring the relative size and overlap of public web search engines. In *Proc. 7th Int. Conf. on World Wide Web*, pages 379–388.

Callan, J. (2000). Distributed information retrieval. In Croft, W. B., editor, *Advances in information retrieval*, volume 7 of *The information retrieval series*, pages 127–150. Kluwer, Dordrecht.

Craswell, N., Bailey, P., and Hawking, D. (2000). Server selection on the world wide web. In *Proc. ACM Int. Conf. on Digital Libraries*, pages 37–46.

Dreilinger, D. and Howe, A. E. (1997). Experiences with selecting search engines using metasearch. *ACM Transactions on Information Systems*, 15(3):195–222.

Elsas, J. L., Arguello, J., Callan, J., and Carbonell, J. G. (2008). Retrieval and feedback models for blog search. In *Proc. 31st Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 347–354.

Gauch, S., Wang, G., and Gomez, M. (1996). ProFusion: Intelligent fusion from multiple, distributed search engines. *Journal of Universal Computer Science*, 2(9):637–649.

Glover, E. J., Lawrence, S., Birmingham, W. P., and Giles, C. L. (1999). Architecture of a metasearch engine that supports user information needs.

In *Proc. 8th ACM Int. Conf. on Information and Knowledge Management*, pages 210–216.

Gordon, M. and Pathak, P. (1999). Finding information on the world wide web: The retrieval effectiveness of search engines. *Information Processing and Management*, 35:141–180.

Gulli, A. and Signorini, A. (2005). The indexable web is more than 11.5 billion pages. In *Proc. 14th Int. Conf. on World Wide Web*, pages 902–903.

Hawking, D., Craswell, N., Bailey, P., and Griffiths, K. (2001). Measuring search engine quality. *Information Retrieval*, 4(1):33–59.

Hawking, D. and Thistlewaite, P. (1999). Methods for information server selection. *ACM Transactions on Information Systems*, 17(1):40–76.

He, B., Patel, M., Zhang, Z., and Chang, K. C.-C. (2007). Accessing the deep web: A survey. *Communications of the ACM*, 50(5):95–101.

Henzinger, M. R., Heydon, A., Mitzenmacher, M., and Najork, M. (2000). On near-uniform URL sampling. In *Proc. 9th Int. Conf. on World Wide Web*, pages 295–308.

Howe, A. E. and Dreilinger, D. (1997). SavvySearch: A meta-search engine that learns which search engines to query. *AI Magazine*, 18(2):19–25.

Kumar, B. T. S. and Pavithra, S. M. (2010). Evaluating the searching capabilities of search engines and metasearch engines: A comparative study. *Annals of Library and Information Studies*, 57:87–97.

Lalmas, M. and Murdock, V., editors (2008). *SIGIR Workshop on Aggregated Search*.

Lawrence, S. and Giles, C. L. (1998). Searching the World Wide Web. *Science*, 280:98–100.

Lawrence, S. and Giles, C. L. (1999). Accessibility of information on the web. *Nature*, 400:107–109.

Lewandowski, D. (2008). A three-year study on the freshness of web search engine databases. *J. Information Science*, 34(6):817–831.

Liu, K.-L., Meng, W., Qui, J., Yu, C., Raghavan, V., Wu, Z., Lu, Y., He, H., and Zhao, H. (2007). AllInOneNews: Development and evaluation of a large-scale news metasearch engine. In *Proc. ACM SIGMOD International Conference on Management of Data*, pages 1017–1028.

Madhavan, J., Afanasiev, L., Antova, L., and Halevy, A. (2009). Harnessing the deep web: Present and future. In *Proc. Biennial Conf. on Innovative Data Systems Research*. arXiv:0909.1785v1 [cs.DB].

Markov, I. (2011). Modeling document scores for distributed information retrieval. In *Proc. 34th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, page 1321. ACM Press.

Meng, W., Yu, C., and Liu, K.-L. (2002). Building efficient and effective metasearch engines. *ACM Computing Surveys*, 34(1):48–89.

Park, S. (1999). User preferences when searching individual and integrated full-text databases. In *Proc. Fourth ACM Conf. on Digital Libraries*, pages 195–203.

Powell, A. L., French, J. C., Callan, J., Connell, M., and Viles, C. L. (2000). The impact of database selection on distributed searching. In *Proc. 23rd*

Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pages 232–239.

Rasolofo, Y., Abbaci, F., and Savoy, J. (2001). Approaches to collection selection and results merging for distributed information retrieval. In *Proc. 10th ACM Int. Conf. on Information and Knowledge Management*, pages 191–198.

Sadeghi, H. (2009). Assessing metasearch engine performance. *Online Information Review*, 33(6):1058–1065.

Selberg, E. and Etzioni, O. (1995). Multi-service search and comparison using the MetaCrawler. In *Proc. 4th Int. Conf. on World Wide Web*.

Shokouhi, M. and Si, L. (2011). Federated search. *Foundations and Trends in Information Retrieval*, 5(1):1–102.

Shu, B. and Kak, S. (1999). A neural network-based intelligent metasearch engine. *Information Sciences*, 120:1–11.

Simeoni, F. (2004). Servicing the federation: The case for metadata harvesting. In *Proc. 8th European Conf. on Research and Advanced Technology for Digital Libraries (ECDL)*, pages 389–399.

Spink, A., Jansen, B. J., Kathuria, V., and Koshman, S. (2006). Overlap among major web search engines. *Internet Research*, 16(4):419–426.

Teevan, J., Alvarado, C., Ackerman, M. S., and Karger, D. R. (2004). The perfect search engine is not enough: A study of orienteering behaviour in directed search. In *Proc. Conf. Human Factors in Computing Systems*, pages 415–422.

Thomas, P. (2008). *Server Characterisation and Selection for Personal Metasearch*. PhD thesis, Australian National University.

Thomas, P. and Hawking, D. (2007). Evaluating sampling methods for uncooperative collections. In *Proc. 30th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 503–510.

Thomas, P. and Shokouhi, M. (2009). SUSHI: Scoring scaled samples for server selection. In *Proc. 32nd Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 419–426.

Wrubel, L. and Schmidt, K. (2007). Usability testing of a metasearch interface: A case study. *College and Research Libraries*, 68(4):292–311.

Xu, J. and Croft, W. B. (1999). Cluster-based language models for distributed retrieval. In *Proc. 22nd Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 254–261.

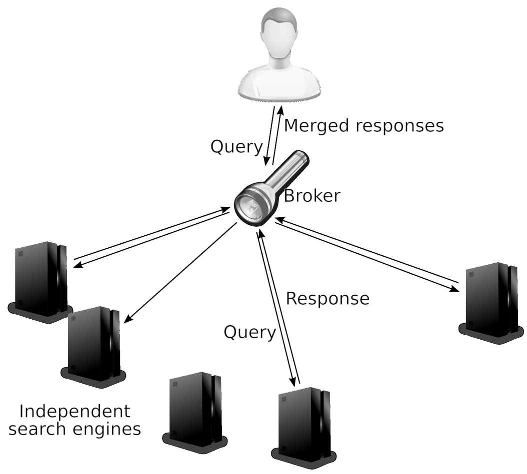

Figure 1: The classic model of distributed IR, considered here. A single *broker* communicates on the user's behalf with one or more search engines from a set. Each search engine runs independently and may, or may not, cooperate by providing special features for the broker. (These are the *cooperative* and *uncooperative* variants, respectively.)

Unlikely	Unproven	Promising
WWW coverage	Ease of use	Deep web
Improved results	Integrating agencies	Personal/enterprise search (private collections)
	Aggregated search	
	Fast-changing collections	

Table 1: Claimed uses for distributed IR, divided three ways: cases where distributed IR is unlikely to help, cases where there is not enough evidence either way, and cases where it is promising.