Collaborative Exploratory Search for Information Filtering and Large-Scale Information Triage

Paul M. Herceg, Timothy B. Allison, Robert S. Belvin, and Evelyne Tzoukermann

The MITRE Corporation

Author Note

Correspondence for this article should be addressed to Paul M. Herceg using the contact information below:

pherceg@mitre.org, 703-983-4539;
tallison@mitre.org, 703-983-2473;
rbelvin@mitre.org, 703-983-1446;
tzoukermann@mitre.org, 703-983-9129;
fax:703-983-1379;
The MITRE Corporation, 7515 Colshire Drive, McLean VA 22102.

This work was supported by Contract W15P7T-13-C-A802, Department of the Army–Aberdeen Proving Ground; Approved for Public Release, Case No. 16-1413. The views, opinions, and/or findings contained in this report are those of The MITRE Corporation and should not be construed as an official government position, policy, or decision, unless designated by other documentation.

Abstract

Modern information seekers face dynamic streams of large-scale heterogeneous data that are both intimidating and overwhelming. They need a strategy to filter this barrage of massive data sets, and to find all of the information responding to their information needs, despite the pressures imposed by schedules and budgets. In this applied research, we present an exploratory search strategy that allows professional information seekers to efficiently and effectively triage all of the data. We demonstrate that exploratory search is particularly useful for information filtering and large-scale information triage, regardless of the language of the data, and regardless of the particular industry, whether finance, medical, business, government, information technology, news, or legal. Our strategy reduces a dauntingly large volume of information into a manageable, high-precision data set, suitable for focused reading. This strategy is interdisciplinary, integrating concepts from information filtering, information triage, and exploratory search. Key aspects include advanced search software, interdisciplinary paired search, asynchronous collaborative search, attention to linguistic phenomena, and aggregated search results in the form of a search matrix or search grid. We present the positive results of a task-oriented evaluation in a real-world setting, discuss these results from a qualitative perspective, and share future research areas.

Keywords: exploratory search, information filtering, information triage, document triage, information retrieval, collaborative information retrieval, collaborative exploratory search, browsing, computer supported cooperative work, paired search, synchronous search, asynchronous search, knowledge discovery, applied research, information seeking, information overload

The challenge of information seeking in a real-world setting can be significantly exacerbated by a number of factors. A knowledge worker can have numerous, multifaceted and detailed information needs (e.g., hundreds, or more). The worker can have a highly dynamic information space. For example, the information space could involve a constant torrent of unstructured data batches that are both high volume and highly heterogeneous (i.e., each being a large set of many different types of files). Such batches might also involve a variety of languages and formats (text, images, audio, and video). Naturally, a business environment adds schedule and budget pressures, necessitating each batch to be searched within a time constraint (e.g., a couple of days). Furthermore, there may be serious consequences to missing any single relevant information object in a batch. Searching in such a space can be overwhelming for humans. This paper contributes an exploratory search strategy, called the *hybrid strategy*, that addresses this challenging setting and enables human knowledge workers to achieve high-precision information filtering and large-scale information triage—a use case rarely addressed in the literature. We describe the implementation of this new strategy in a real-world setting using a minimally-intrusive survey instrument.

The *hybrid strategy* sequences exploratory search and directed browsing strategies described in the existing search literature. Marchionini (1997) divided human information seeking using an information system into two major categories—*analytical search strategies* and *browsing strategies*—which are the two extremes of a continuum (pp. 8-9). Analytical search involves pre-determining search terms. In Blair and Maron (1985), legal professionals planned out terms, queried a search system, and yielded relevant documents. The professionals believed the yield was 75 percent of the relevant documents, but it was only 20 percent. Users rely on analytical search when their information seeking task is time-constrained. However, alone, analytical search misses a majority of the relevant material.

Therefore, users turn to the opposing strategy, *browsing*, which involves comprehensively "scanning, jumping, and navigating" file content, for each file of a large data set; hereafter, we discuss *directed browsing*, which is "scanning, jumping, and navigating" with the purpose of finding specific information (Marchionini, 1997, pp. 106, 158). This approach gives the information seeker comfort that he/she is finding every information object responding to the information need. However, browsing, too, falls short because human fatigue is a real physiological and psychological barrier; furthermore, browsing is significantly less efficient than search, and browsing can only cover a fraction of a large information space (Marchionini, 1997, pp. 117-118).

Marchionini (2006) used the term *exploratory search* to describe the combination of an analytical search strategy and a browsing strategy, which overcomes the limitations of either extreme. According to his concept, exploratory search accommodates information seekers who are unfamiliar with the information objects in a search space, and who have yet to obtain the knowledge needed to formulate effective queries in that space. Therefore, exploratory search is useful where there is both a multiplicity of complex information needs and a large and dynamic information space.

The design of the *hybrid strategy* allows an information seeker to reduce a dauntingly large volume of information into a manageable, high-precision data set, suitable for focused reading. The strategy enables information seekers to efficiently and effectively "find the needle in the haystack", or numerous needles in the haystack. The strategy developed organically during collaboration with a real-world group of information seekers, a group characterized by the complex information needs and information space outlined in the first paragraph of this paper. This group was skilled at applying a directed browsing strategy within a time constraint, which we later characterize as information triage. The conventional task of each browsing expert involved applying both linear and non-linear browsing strategies—selectively scanning, jumping, and navigating the content of each file within a batch of files. When an expert recognized an information object (file content) that responded to the information need, he/she sequestered the file. Each expert accumulated the sequestered files and sent them as findings to a separate follow-on group for focused reading. The hybrid strategy inserted an exploratory search activity prior to this directed browsing activity, augmenting and amplifying their time-sensitive information seeking work, in a manner that was compatible with their conventional work style. We hypothesized that the hybrid strategy would result in a large positive effect on the efficiency and effectiveness of the browsing experts. We used a survey instrument to measure this positive effect. The hybrid strategy did result in strongly positive Likert item responses for utility, speed, and quality.

Regarding our use of the term *large* to describe data sets, we deliberately set no upper bound. We mean an unwieldy size that poses a challenge for human information seekers, minimally gigabytes. In this study, we observed the processing of hundreds of gigabytes, but *large* data could be magnitudes greater in size.

Technology adoption and implementation (i.e., putting the technology into use) was a chief consideration for our applied research (Rogers, 2003, pp. 179-180). The hybrid strategy was welcomed by the browsing experts because it increased two attributes of technology known to amplify the rate of adoption: (a) the relative advantage over the incumbent method, and (b) the compatibility with the values and needs of users (Rogers, 2003, p. 222).

Our contribution is the *hybrid strategy* that we integrated into a real-world setting, supported by qualitative results. Furthermore, this strategy is fully reproducible.

Related Work

Belkin and Croft (1992) differentiated *information filtering* from information retrieval, text categorization, and information extraction, as follows: filtering systems distill, or cull, large volumes of data (gigabytes/terabytes), where the unstructured or semi-structured input (text, image, audio, video) arrives as a data stream, rather than as a static corpus. The authors explained that information filtering seeks to match incoming documents against profiles (queries), each of which represents an individual’s/group’s information need, and deliver to users the most relevant documents. They added that profiles have long-term, enduring value, but can change over time. Information filtering is a strategy that information seekers use to deal with information overload (Case, 2012, p. 116). For years, NIST TREC hosted a series of tracks that evaluated information filtering systems (Voorhees & Harman, 2005, pp. 99-101). Büttcher, Clarke, and Cormack (2010, p. 313) discuss *batch filtering*: incoming documents accumulate into a corpus/batch, an IR system uses each profile to retrieve and rank documents, ranked lists are delivered to users, and the entire process iterates. To the literature on information filtering, our paper contributes a human-in-the-loop approach for large-volume, high-precision batch filtering.

Marshall and Shipman (1997) defined *information triage*, or document triage, as the activity of information seekers sorting through information resources under some time constraint. Increasingly, triage is performed against large-scale, unstructured or semi-structured document collections, differentiating relevant documents from irrelevant ones. Buchanan and Loizides (2007) and Loizides and Buchanan (2009) mention that (a) the scientific understanding of document triage is limited, (b) the research on triage is fragmented, and (c) triage has attracted little attention. Existing studies of triage include text classifiers for prioritization (Macskassy & Provost, 2001), display configurations (Bae et al., 2005), user activity logging (Badi et al., 2006; Bae et al., 2010), paper versus electronic documents (Buchanan & Loizides, 2007), enhancing section headings (Buchanan & Owen, 2008), visual search patterns (Loizides & Buchanan, 2009), and tag clouds (Maiya, Thompson, Loaiza-Lemos, & Rolfe, 2013). Aside from the latter, studies use uniform document formats and small information spaces (e.g., 200 documents). However, real-world information triage involves heterogeneous digital document formats (Allison & Herceg, 2015) and dramatic reduction/culling of large data volumes. To the information triage literature, we contribute an information triage approach that (a) serializes two distinct and differing passes of triage, and (b) addresses data volumes far beyond typical studies.

According to Marchionini (2006), exploratory search involves interactive information retrieval and a number of learning and investigation activities; his list and discussion of these latter activities includes lookup, comprehension, interpretation, comparison, aggregation, integration, accretion, reformulation, analysis, exclusion, synthesis, evaluation (scanning/viewing/judging), discovery (serendipity), and transformation (pp. 42-43). White and Roth (2009) added that exploratory search (a) is conducted over variable time spans, (b) is “open-ended, persistent, and multifaceted”, (c) amplifies the intelligence of the information seeker, and (d) combines both browsing and focused searching (pp. 10, 21-22). *Collaborative* exploratory search involves multiple people, with the same information need, working together, either asynchronously or synchronously (Pickens and Golovchinsky, 2007). Kelly and Payne (2013) studied the division of labor (DoL) among collaborative searchers and found user interface, algorithmic, and role-based DoL approaches. “Prior studies of role-based DoL have only examined concurrent search scenarios [synchronous, paired work], and little is known about the benefit of roles for asynchronous work” (p. 3). To the exploratory search literature, we contribute a study involving both synchronous and asynchronous collaborative exploratory search.

Morris, Morris, and Venolia (2008) showed that users suspend and resume searches, and that the gap between suspension and resumption varies widely, from minutes to years. Venolia (2008) (a) integrated search queries and results into a single search grid/matrix, (b) captured, stored, and presented search query and result history, and (c) allowed each query to be refined and repeated over time. Our work employed similar techniques, and used a different grid/matrix to facilitate the asynchronous work of the hybrid strategy (DoL).

Macskassy and Provost (2001) pointed out that timely reports from vast information sources (filtering and large-scale triage) are important to a variety of professionals: considering business news alone, the users include “financial analysts, attorneys, business-school professors, market makers, portfolio managers, reporters” (p. 318). Additional applications are business intelligence, e-discovery early case assessment, auditing, and IT helpdesk. Pirolli and Card (2005) presented a model of the work process of analysts. Our integration of information filtering, information triage, and exploratory search contributes to the initial steps of this model.

Method

The hybrid strategy serialized exploratory search and directed browsing, each yielding an independent pass of information triage. For 5 months, this strategy processed several large, incoming batches of highly heterogeneous media (hundreds of types). We used questionnaires (below) to gather quantitative and qualitative data. Because we applied only full-text search, we constrained the evaluation to the text-containing file types.

In the conventional task, a group of browsing experts processed incoming batches of data, consisting of highly heterogeneous media, striving to not miss any single relevant information object in each batch. Files were primarily in a foreign language, but also contained English and other languages. Each browsing expert possessed an understanding of (a) the list of enduring information needs across batches, and (b) the list of particular information needs pertinent to each batch, which could be looked up. There were two groups of browsing experts: Group 1 was skilled in the foreign language and English, and worked under a two-day timespan; Group 2 had expertise in English only, used machine translation to access foreign language content, and had a much more variable timespan (several days). This conventional setting existed at the beginning of our study and is depicted in Figure 1(a).

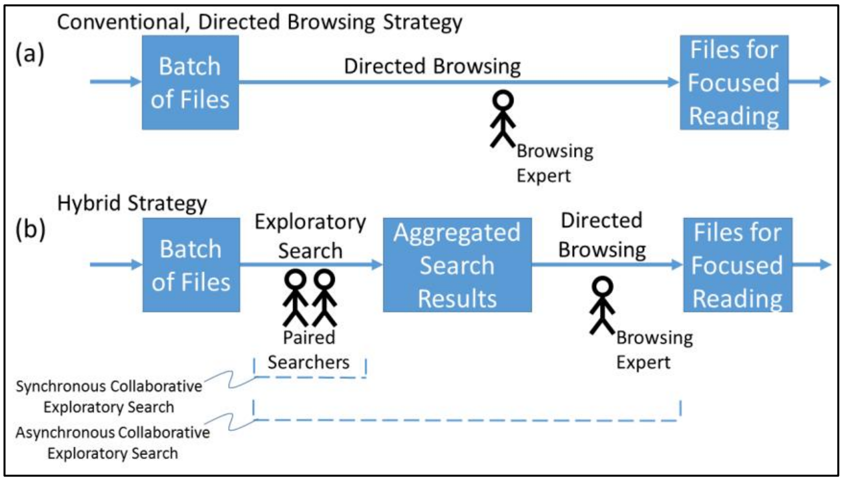

Conventional, Directed Browsing Strategy
(a)
Batch of Files
Directed Browsing
Files for Focused Reading
Browsing Expert
Hybrid Strategy
(b)
Batch of Files
Exploratory Search
Aggregated Search Results
Directed Browsing
Files for Focused Reading
Paired Searchers
Browsing Expert
Synchronous Collaborative Exploratory Search
Asynchronous Collaborative Exploratory Search

Figure 1. The conventional directed browsing strategy, and the hybrid strategy.

The hybrid strategy extended this conventional design by adding exploratory search, depicted in Figure 1(b). Interdisciplinary paired searchers performed synchronous collaborative exploratory search on arriving batches, unaware of the content of each. On arrival of a batch they (a) consulted with the designated browsing expert to elicit the information needs (e.g., terms), and (b) looked up the information needs for the batch in the same manner as the browsing expert did. This understanding of information needs steered the investigative search activity. The paired searchers accumulated sets of search results, each of which was correlated with a specific information need. Accumulated search results were aggregated for the browsing expert into a single knowledge product, the *aggregated search results*. The expert analyzed these results, used them to prioritize files, performed directed browsing of indicated files, and sequestered files for focused reading. Each expert first reviewed the files appearing in the aggregated search results, and, only afterward, if time allowed, browsed other files in the batch. At the close of his/her browsing activity, each expert completed a questionnaire (below) indicating the value of the aggregated search results to his/her browsing activity. Only Group 1 and Group 2 experts responded to the questionnaire. Together, the paired searchers and the browsing expert effected *asynchronous collaborative exploratory search*, as shown in Figure 1(b).

The browsing strategy mandated that aggregated search results carry high precision; the browsing experts had no tolerance for new browsing aids that pointed to false positives, i.e., files that did not respond to the information needs.

Interdisciplinary Paired Searchers

The hybrid strategy required interdisciplinary paired searchers. The first was a Ph.D. *computational linguist,* an English speaker with no expertise in the foreign language. The second searcher was a *domain expert,* with four skills: (a) ability to speak, read, and type in the foreign language, (b) fluency in English, (c) an expert understanding of the information needs and the domain of the incoming batches, and (d) expertise in the directed browsing task performed by the group of browsing experts. We co-opted two senior browsing experts for this role: one for the first half of the study, and another for the second.

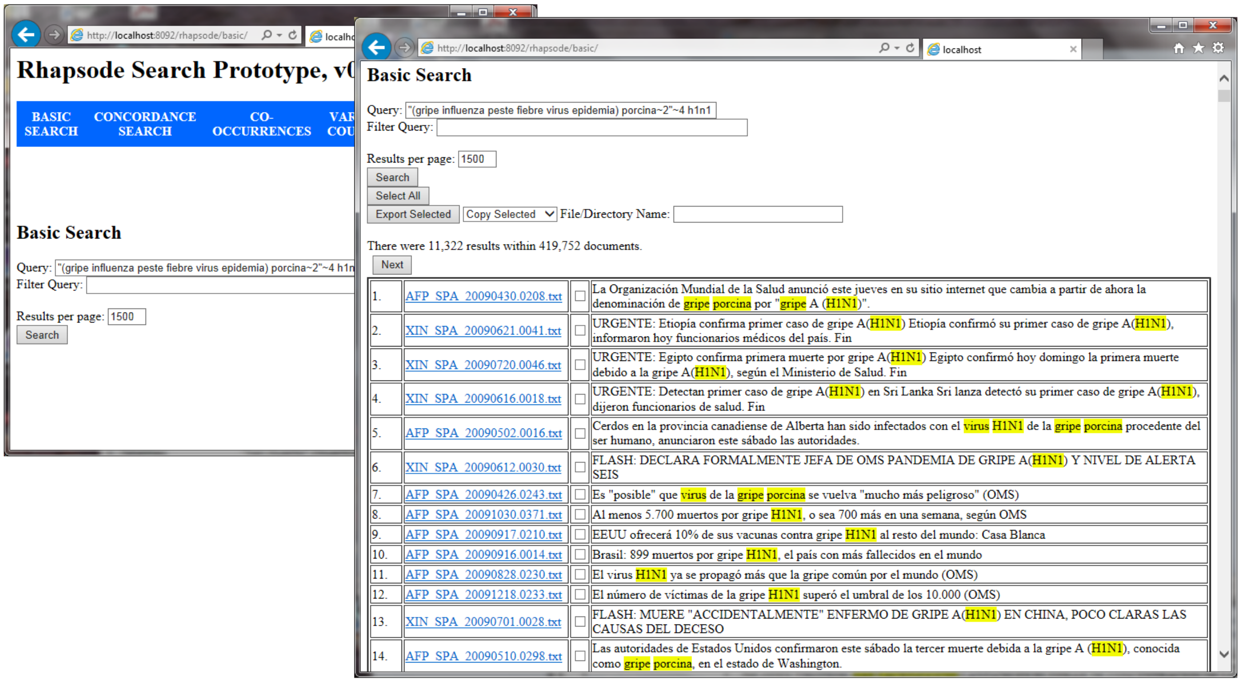
The hybrid strategy was implemented with MITRE Rhapsode v.0.21 advanced search (below). The paired searchers used Rhapsode and the computational linguist was the lead user.

Exploratory Search Software

Rhapsode is a search application built with Apache Lucene for information retrieval (http://lucene.apache.org), and Apache Tika for text extraction (http://tika.apache.org). Rhapsode was instrumental to the implementation of the hybrid strategy because of its text extraction, query operators, multilingual handling, and document viewing features. Rhapsode offered a combination of features that was not available in Web or commercial search engines:

- Ability to process gigabytes of data on a single business class PC (i.e., lightweight)
- Lucene query operators (including wildcards, fuzzy term search, i.e., character variations)
- An extended set of query operators:
  - Wildcard and fuzzy term search within phrases
  - Proximity search of phrases
  - Specification of “in order” or “any order” for proximity searches
  - Search for fuzzy terms with a number of character variations beyond the standard Lucene default of 2
- Conscientious use of Lucene analyzers, including robust multilingual tokenization and Unicode normalization, to improve recall
- Ability to display long lists of search results (e.g., 1500)
- Hit-highlighting in both search results and documents selected from search results
- Term co-occurrence statistics (i.e., collocations)
- Ability to display a long list of search results as a concordance
- Index statistics on term matches to Lucene multiterm queries (wildcard/fuzzy/regex) for enabling high precision retrieval

The combination of Lucene query operators and the extended set of phrasal query operators in Rhapsode allowed the paired searchers to retrieve minute details from a batch of files (i.e., “the needles in the haystack”). The wide set of operators in Rhapsode was useful for investigating the information space and discovering collocations, and variants such as syntactic, morphological, morphosyntactic, lexical, and orthographic (including misspellings). Figures 2 and 3 show some screenshots of Rhapsode.



http://localhost:8092/rhapsode/basic/

Rhapsode Search Prototype, v0

Basic Search

Query: "(gripe influenza peste fiebre virus epidemia) porcina~2"~4 h1n
Filter Query:

Results per page: 1500

http://localhost:8092/rhapsode/basic/

Basic Search

Query: "(gripe influenza peste fiebre virus epidemia) porcina~2"~4 h1n1
Filter Query:

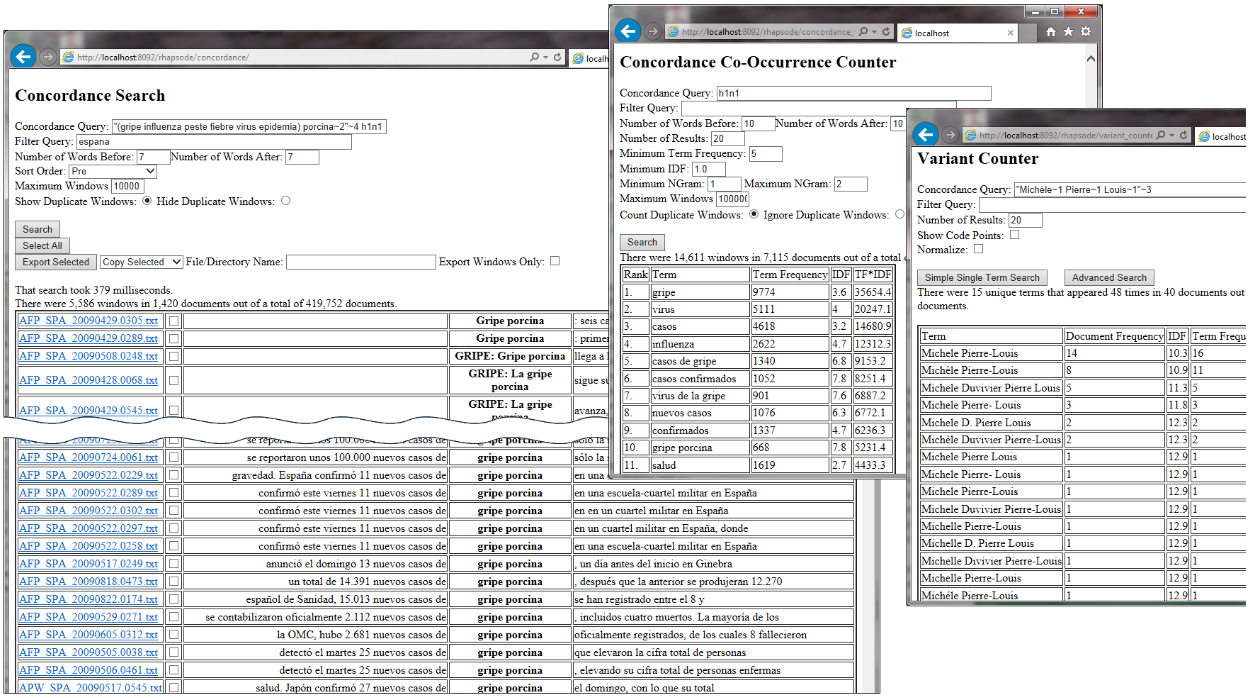
Results per page: 1500

File/Directory Name:

There were 11,322 results within 419,752 documents.

1.	AFP_SPA_20090430.0208.txt	□	La Organización Mundial de la Salud anunció este jueves en su sitio internet que cambia a partir de ahora la denominación de gripe porcina por "gripe A (H1N1)".
2.	XIN_SPA_20090621.0041.txt	□	URGENTE: Etiopía confirma primer caso de gripe A(H1N1) Etiopía confirmó su primer caso de gripe A(H1N1), informaron hoy funcionarios médicos del país. Fin
3.	XIN_SPA_20090720.0046.txt	□	URGENTE: Egipto confirma primera muerte por gripe A(H1N1) Egipto confirmó hoy domingo la primera muerte debido a la gripe A(H1N1), según el Ministerio de Salud. Fin
4.	XIN_SPA_20090616.0018.txt	□	URGENTE: Detectan primer caso de gripe A(H1N1) en Sri Lanka Sri lanza detectó su primer caso de gripe A(H1N1), dijeron funcionarios de salud. Fin
5.	AFP_SPA_20090502.0016.txt	□	Cerdos en la provincia canadiense de Alberta han sido infectados con el virus H1N1 de la gripe porcina procedente del ser humano, anunciaron este sábado las autoridades.
6.	XIN_SPA_20090612.0030.txt	□	FLASH: DECLARA FORMALMENTE JEFA DE OMS PANDEMIA DE GRIPE A(H1N1) Y NIVEL DE ALERTA SEIS
7.	AFP_SPA_20090426.0243.txt	□	Es "posible" que virus de la gripe porcina se vuelva "mucho más peligroso" (OMS)
8.	AFP_SPA_20091030.0371.txt	□	Al menos 5.700 muertos por gripe H1N1, o sea 700 más en una semana, según OMS
9.	AFP_SPA_20090917.0210.txt	□	EEUU ofrecerá 10% de sus vacunas contra gripe H1N1 al resto del mundo: Casa Blanca
10.	AFP_SPA_20090916.0014.txt	□	Brasil: 899 muertos por gripe H1N1, el país con más fallecidos en el mundo
11.	AFP_SPA_20090828.0230.txt	□	El virus H1N1 ya se propagó más que la gripe común por el mundo (OMS)
12.	AFP_SPA_20091218.0233.txt	□	El número de víctimas de la gripe H1N1 superó el umbral de los 10.000 (OMS)
13.	XIN_SPA_20090701.0028.txt	□	FLASH: MUERE "ACCIDENTALMENTE" ENFERMO DE GRIPE A(H1N1) EN CHINA, POCO CLARAS LAS CAUSAS DEL DECESO
14.	AFP_SPA_20090510.0298.txt	□	Las autoridades de Estados Unidos confirmaron este sábado la tercer muerte debida a la gripe A (H1N1), conocida como gripe porcina, en el estado de Washington.

Figure 2. Rhapsode Basic Search interface. This presents standard snippets of the first three times a term appears in a document.



http://localhost:8092/rhapsode/concordance/

Concordance Search

Concordance Query: "(gripe influenza peste fiebre virus epidemia) porcina~2"~4 h1n1
Filter Query: espana
Number of Words Before: 7 Number of Words After: 7
Sort Order: Pre
Maximum Windows 10000
Show Duplicate Windows: ● Hide Duplicate Windows: ○

File/Directory Name: Export Windows Only: □

That search took 379 milliseconds.
There were 5,586 windows in 1,420 documents out of a total of 419,752 documents.

AFP_SPA_20090429.0305.txt	□		**Gripe porcina**	: seis ca
AFP_SPA_20090429.0289.txt	□		**Gripe porcina**	: primer
AFP_SPA_20090508.0248.txt	□		**GRIPE: Gripe porcina**	llega a l
AFP_SPA_20090428.0068.txt	□		**GRIPE: La gripe porcina**	sigue su
AFP_SPA_20090429.0545.txt	□		**GRIPE: La gripe porcina**	avanza,
AFP_SPA_20090724.0061.txt	□	se reportaron unos 100.000 nuevos casos de	**gripe porcina**	sólo la
AFP_SPA_20090522.0229.txt	□	gravedad. España confirmó 11 nuevos casos de	**gripe porcina**	en una
AFP_SPA_20090522.0289.txt	□	confirmó este viernes 11 nuevos casos de	**gripe porcina**	en una escuela-cuartel militar en España
AFP_SPA_20090522.0302.txt	□	confirmó este viernes 11 nuevos casos de	**gripe porcina**	en en un cuartel militar en España
AFP_SPA_20090522.0297.txt	□	confirmó este viernes 11 nuevos casos de	**gripe porcina**	en un cuartel militar en España, donde
AFP_SPA_20090522.0258.txt	□	confirmó este viernes 11 nuevos casos de	**gripe porcina**	en una escuela-cuartel militar en España
AFP_SPA_20090517.0249.txt	□	anunció el domingo 13 nuevos casos de	**gripe porcina**	, un día antes del inicio en Ginebra
AFP_SPA_20090818.0473.txt	□	un total de 14.391 nuevos casos de	**gripe porcina**	, después que la anterior se produjeran 12.270
AFP_SPA_20090822.0174.txt	□	español de Sanidad, 15.013 nuevos casos de	**gripe porcina**	se han registrado entre el 8 y
AFP_SPA_20090529.0271.txt	□	se contabilizaron oficialmente 2.112 nuevos casos de	**gripe porcina**	, incluidos cuatro muertos. La mayoría de los
AFP_SPA_20090605.0312.txt	□	la OMC, hubo 2.681 nuevos casos de	**gripe porcina**	oficialmente registrados, de los cuales 8 fallecieron
AFP_SPA_20090505.0038.txt	□	detectó el martes 25 nuevos casos de	**gripe porcina**	que elevaron la cifra total de personas
AFP_SPA_20090506.0461.txt	□	detectó el martes 25 nuevos casos de	**gripe porcina**	, elevando su cifra total de personas enfermas
APW_SPA_20090517.0545.txt	□	salud. Japón confirmó 27 nuevos casos de	**gripe porcina**	el domingo, con lo que su total

http://localhost:8092/rhapsode/concordance_

Concordance Co-Occurrence Counter

Concordance Query: h1n1
Filter Query:
Number of Words Before: 10 Number of Words After: 10
Number of Results: 20
Minimum Term Frequency: 5
Minimum IDF: 1.0
Minimum NGram: 1 Maximum NGram: 2
Maximum Windows 10000(
Count Duplicate Windows: ● Ignore Duplicate Windows: ○

There were 14,611 windows in 7,115 documents out of a total

Rank	Term	Term Frequency	IDF	TF*IDF
1.	gripe	9774	3.6	35654.4
2.	virus	5111	4	20247.1
3.	casos	4618	3.2	14680.9
4.	influenza	2622	4.7	12312.3
5.	casos de gripe	1340	6.8	9153.2
6.	casos confirmados	1052	7.8	8251.4
7.	virus de la gripe	901	7.6	6887.2
8.	nuevos casos	1076	6.3	6772.1
9.	confirmados	1337	4.7	6236.3
10.	gripe porcina	668	7.8	5231.4
11.	salud	1619	2.7	4433.3

http://localhost:8092/rhapsode/variant_count

Variant Counter

Concordance Query: "Michèle~1 Pierre~1 Louis~1"~3
Filter Query:
Number of Results: 20
Show Code Points: □
Normalize: □

There were 15 unique terms that appeared 48 times in 40 documents out documents.

Term	Document Frequency	IDF	Term Frequ
Michele Pierre-Louis	14	10.3	16
Michèle Pierre-Louis	8	10.9	11
Michele Duvivier Pierre Louis	5	11.3	5
Michele Pierre- Louis	3	11.8	3
Michele D. Pierre Louis	2	12.3	2
Michèle Duvivier Pierre-Louis	2	12.3	2
Michele Pierre Louis	1	12.9	1
Michele Pierre- Louis	1	12.9	1
Michele Pierre-Louis	1	12.9	1
Michele Duvivier Pierre-Louis	1	12.9	1
Michelle Pierre-Louis	1	12.9	1
Michelle D. Pierre Louis	1	12.9	1
Michelle Divivier Pierre-Louis	1	12.9	1
Michelle Pierre-Louis	1	12.9	1
Michéle Pierre-Louis	1	12.9	1

Figure 3. Rhapsode Concordance search, Co-Occurrence Counter and Variant Counter. The Concordance shows every term occurrence in a batch, allowing review, in context, within and across documents. The Co-Occurrence Counter calculates statistics to recommend terms. The Variant Counter shows the spelling variations found in the batch for a given query.

Reporting the details for this study was challenging because the data and detailed use case were subject to privacy. Thus, Figure 2 and Figure 3, and forthcoming figures 4, 5, 9, and 10, depict modified data. These figures show a biomedical research situation, where paired searchers and browsing experts are seeking detailed information related to disease outbreak events, and are processing dynamic reports in Spanish. The depicted data is from Mendonça et al. (2011), a corpus of Spanish language news in a standard file format. Recall that the actual incoming batches consisted of highly heterogeneous and unstructured data.

Paired Searching

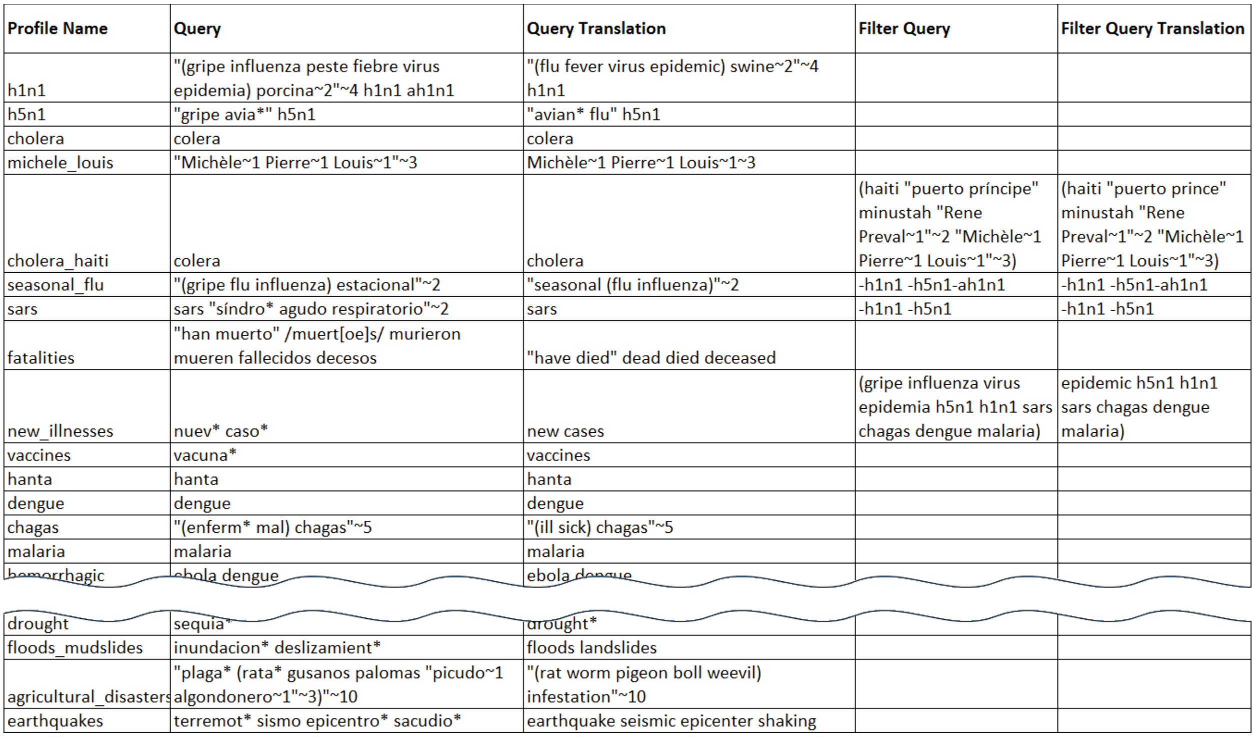
The paired searchers worked synchronously at a single PC, using Rhapsode to formulate and refine profile queries that yielded relevant files, and that excluded irrelevant files. The task was as follows: (a) iteratively develop and refine a set of profile queries for the batch, and (b) save the profile search results for each profile query that yielded a response to the information needs.

By taking full advantage of the Rhapsode operators, and becoming intimately acquainted with the relevant content of each batch, the paired searchers optimized the yield of information responding to each information need. Each iteration that yielded new discoveries (collocations and linguistic variants) resulted in the refinement of a given profile and/or the formulation of new profiles, informed by the discoveries. Iterations continued until the searchers were satisfied that they had elicited sufficient content.

In order to meet the high-precision requirement, searchers used three strategies: (a) refine the profile terms to be more discriminating, (b) specify exclusion arguments, and (c) specify the rank, in the search results, where precision was approximately 90%, henceforth called a *cutoff rank*. Each list of profile search results was saved with a cutoff rank. Only returns above and including the cutoff rank were included in aggregated search results.

Search sessions were suspended and resumed, as needed, over the course of hours or days. Individual profile queries were saved during their formulation. Lists of profile search results were saved for later aggregation. Each individual list of profile search results was saved as an individual file, organized by profile name, designating the determined cutoff rank.

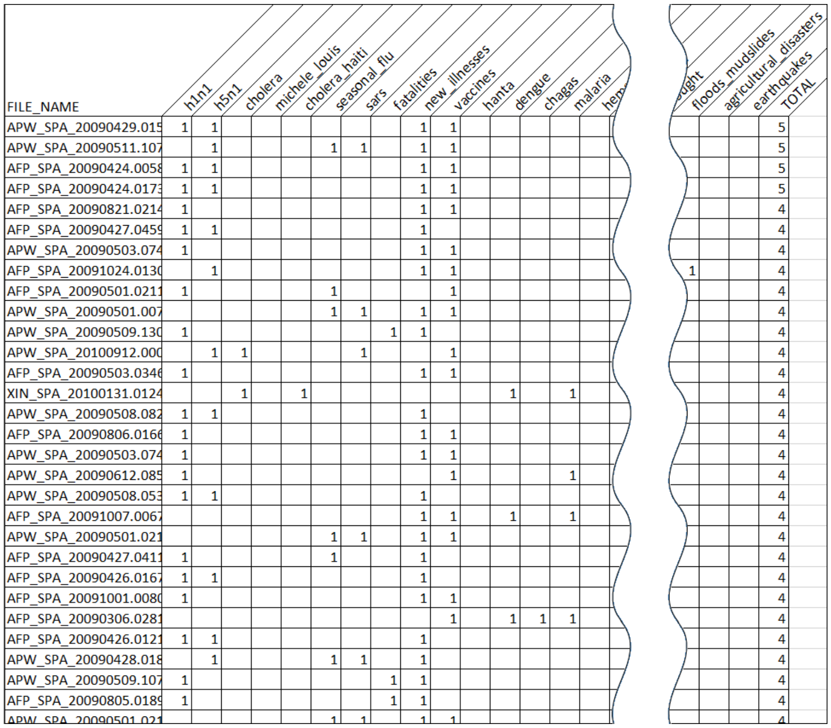
Each profile was saved by adding it to the *master profile query dictionary*. An example is Figure 4, an Excel worksheet with five fields. The query (column 2) captures all aforementioned linguistic variants corresponding to a given concept/facet, in Rhapsode Lucene-like syntax. In the first row, the profile matches multiple variants of H1N1 (“gripe”, “influenza”, “peste”, “fiebre”, “virus”, and “epidemia”) in proximity with a word for swine (“porcina”), or the literal designators “h1n1” and “ah1n1”. Further matches include morphological/spelling variants and misspellings, e.g., “porcina”, “*procina”, or “porcinos” as in “virus en cerdos [porcinos]”. Moreover, the profile will match any file containing “porcina” with 1 or 2 character edits (insertions, deletions, or transpositions) within 4 words of any one of the 6 terms for illness. The permutations of these variants can yield several hundred different literal queries.



Profile Name	Query	Query Translation	Filter Query	Filter Query Translation
h1n1	"(gripe influenza peste fiebre virus epidemia) porcina~2"~4 h1n1 ah1n1	"(flu fever virus epidemic) swine~2"~4 h1n1		
h5n1	"gripe avia*" h5n1	"avian* flu" h5n1		
cholera	colera	colera		
michele_louis	"Michèle~1 Pierre~1 Louis~1"~3	Michèle~1 Pierre~1 Louis~1~3		
cholera_haiti	colera	cholera	(haiti "puerto príncipe" minustah "Rene Preval~1"~2 "Michèle~1 Pierre~1 Louis~1"~3)	(haiti "puerto prince" minustah "Rene Preval~1"~2 "Michèle~1 Pierre~1 Louis~1"~3)
seasonal_flu	"(gripe flu influenza) estacional"~2	"seasonal (flu influenza)"~2	-h1n1 -h5n1-ah1n1	-h1n1 -h5n1-ah1n1
sars	sars "síndro* agudo respiratorio"~2	sars	-h1n1 -h5n1	-h1n1 -h5n1
fatalities	"han muerto" /muert[oe]s/ murieron mueren fallecidos decesos	"have died" dead died deceased		
new_illnesses	nuev* caso*	new cases	(gripe influenza virus epidemia h5n1 h1n1 sars chagas dengue malaria)	epidemic h5n1 h1n1 sars chagas dengue malaria)
vaccines	vacuna*	vaccines		
hanta	hanta	hanta		
dengue	dengue	dengue		
chagas	"(enferm* mal) chagas"~5	"(ill sick) chagas"~5		
malaria	malaria	malaria		
hemorrhagic	ebola dengue	ebola dengue		
drought	sequia*	drought*		
floods_mudslides	inundacion* deslizamient*	floods landslides		
agricultural_disasters	"plaga* (rata* gusanos palomas "picudo~1 algondonero~1"~3)"~10	"(rat worm pigeon boll weevil) infestation"~10		
earthquakes	terremot* sismo epicentro* sacudio*	earthquake seismic epicenter shaking		

Figure 4. Example master profile query dictionary.

At the close of the exploratory search session for a given batch, the full set of individual profile search results was merged into *aggregated search results* (a matrix), delivered to the browsing expert for information triage (Figure 5). The matrix had two dimensions (files and profiles) with a value (“1”). A “1” indicated the presence of a profile search result. Because a file could be present on one or more of the profile search results, rows were merged such that each file was listed only once and all search results were preserved. The ones were summed to the right for a global score, on which files were sorted, in descending order (prioritization). The more matching profiles for a file, the higher the file appeared in the matrix. Figure 6 shows the products of each exploratory search session.



FILE_NAME	h1n1	h5n1	cholera	michele_louis	cholera_haiti	seasonal_flu	sars	fatalities	new_illnesses	vaccines	hanta	dengue	chagas	malaria	hem	ught	floods_mudslides	agricultural_disasters	earthquakes	TOTAL
APW_SPA_20090429.015	1	1							1	1										5
APW_SPA_20090511.107		1				1	1		1	1										5
AFP_SPA_20090424.0058	1	1							1	1										5
AFP_SPA_20090424.0173	1	1							1	1										5
AFP_SPA_20090821.0214	1								1	1										4
AFP_SPA_20090427.0459	1	1							1											4
APW_SPA_20090503.074	1								1	1										4
AFP_SPA_20091024.0130		1							1	1							1			4
AFP_SPA_20090501.0211	1					1				1										4
APW_SPA_20090501.007						1	1		1	1										4
APW_SPA_20090509.130	1							1	1											4
APW_SPA_20100912.000		1	1				1			1										4
AFP_SPA_20090503.0346	1								1	1										4
XIN_SPA_20100131.0124			1		1							1		1						4
APW_SPA_20090508.082	1	1							1											4
AFP_SPA_20090806.0166	1								1	1										4
APW_SPA_20090503.074	1								1	1										4
APW_SPA_20090612.085	1									1				1						4
APW_SPA_20090508.053	1	1							1											4
AFP_SPA_20091007.0067									1	1		1		1						4
APW_SPA_20090501.021						1	1		1	1										4
AFP_SPA_20090427.0411	1					1			1											4
AFP_SPA_20090426.0167	1	1							1											4
AFP_SPA_20091001.0080	1								1	1										4
AFP_SPA_20090306.0281										1		1	1	1						4
AFP_SPA_20090426.0121	1	1							1											4
APW_SPA_20090428.018		1				1	1		1											4
APW_SPA_20090509.107	1							1	1											4
AFP_SPA_20090805.0189	1							1	1											4
APW_SPA_20090501.021						1	1		1	1										4

Figure 5. Example of aggregated search results.

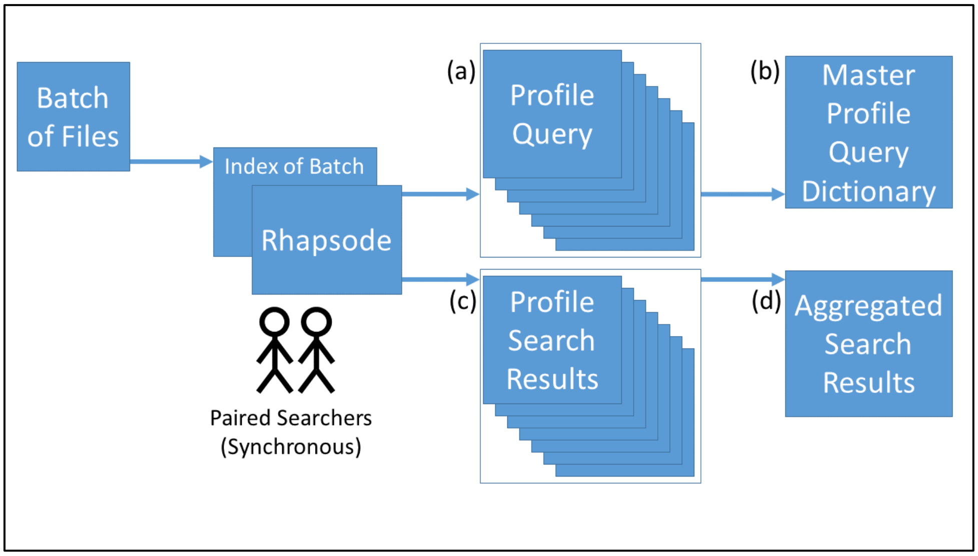

Batch of Files
Index of Batch
Rhapsode
Paired Searchers (Synchronous)
(a)
Profile Query
(b)
Master Profile Query Dictionary
(c)
Profile Search Results
(d)
Aggregated Search Results

Figure 6. The products of each exploratory search session:
(a) profile queries, (b) master profile query dictionary,
(c) lists of profile search results, (d) aggregated search results.

An Excel spreadsheet for the aggregated search results allowed the recipients to easily sort the matrix by any combination of profiles (columns). The results were both static (the information in the Excel spreadsheet did not change), and dynamic (a receiving browsing expert could easily sort the results).

The delivery included the aggregated search results, a relevant subset of the master profile query dictionary, and all Tika text-extracted renditions of the original files.

Due to the time restrictions, we imposed two heuristics on the paired searchers. First, *the duration for formulating a profile query should be one hour, on average*: it was difficult for them (a) to know when to cease the development or refinement of a given profile query, and (b) to judge when the query was good enough. Second, *the duration of an exploratory search session for a batch should be two days, on average*: it was challenging for them to know when to cease developing additional profile queries (for large batches, more exploratory search effort yielded more discoveries).

Measurement

Table 1 and Table 2 characterize the questionnaires we used. Each included 5-point Likert items and reflected the different skill sets of Group 1 and Group 2. The main differences resided in the language skill and the allotted task time.

Table 1

Questionnaire detail (Browsing Expert Group 1)

Question	Response Type	Aspect
The [Aggregated Search Results] helped me perform my job.	Likert	Utility
Based on my job experience, the [Aggregated Search Results] helped me find…files faster (i.e., faster than I would have without the [Aggregated Search Results].)	Likert	Speed
Based on my job experience, the [Aggregated Search Results] helped me to not miss …files (i.e., the report helped me to not inadvertently overlook pertinent files.)	Likert	Quality
How many hours (or days) of effort did the [Aggregated Search Results] save you?	Duration	Cost-Benefit
Comments	Free-form text	Other

Table 2

Questionnaire detail (Browsing Expert Group 2)

Question	Response Type	Aspect
Based on my job experience, the information/file(s) I received from the [paired searchers] helped me perform my job.	Likert	Utility
Based on my job experience, the information/file(s) I received from the [paired searchers] would have been difficult to find without this help. Reasons include one of the following: • Time constraints for reviewing this media • The procedure used to skim documents (i.e., document review) • Other	Likert	Discovery
How many hours (or days) of effort did the information/file(s) save you?	Duration	Cost-Benefit
Comments	Free-form text	Other

Results

In this section, we report both quantitative results and qualitative observations. The qualitative observations are structured by five topic areas.

The hybrid strategy was applied to incoming batches for 5 months. The storage size across all batches was 611.95GB, from which 5,619,027 files were processed. A total of 10,132 files, 409 profiles, and 12,329 search results, appeared across all the aggregated search results. During the study the master profile query dictionary grew from 153 to 1034 queries.

Completed questionnaires were requested for each set of aggregated search results. Among the numerous batches processed, we collected only 30 questionnaires. Each bar of Figure

7 and Figure 8 represents a single Likert item response. Group 1 ratings showed 41% “Agree” and 39% “Strongly Agree”, resulting in 80% combined. Ratings increased over time to a consistently high level: the latter half of all ratings collected were solely “Agree” and “Strongly Agree”. In contrast, Group 2 ratings were high, 12% “Agree” and 88% “Strongly Agree”, resulting in 100% combined.

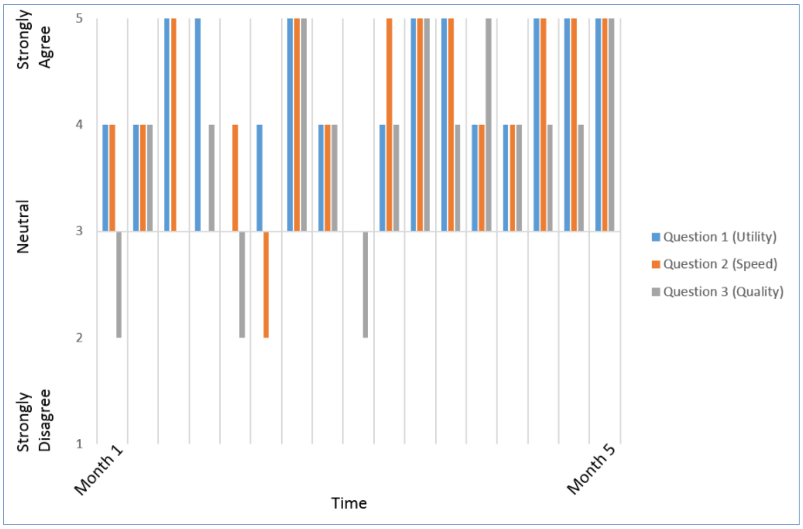


Figure 7. Likert item responses from Browsing Expert Group 1.

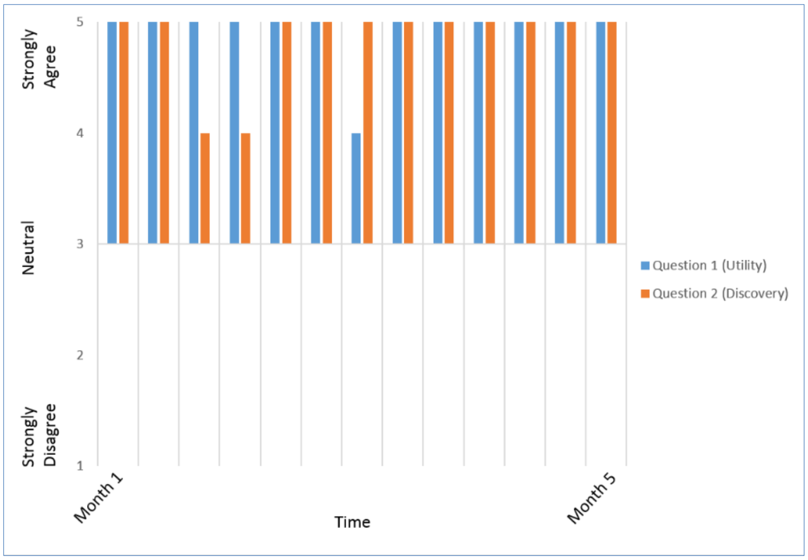


Figure 8. Likert item responses from Browsing Expert Group 2.

The following Table 3 and Table 4 report time/labor savings estimates. Table 3 shows the Group 1 reports of hours saved for 17 batches. Savings averaged 5 hours, ranging from 0 to 16 hours, 4 responses of which were 10 to 16 hours. Recall that this group worked according to a two-day timespan. The experts reported zero savings when a batch yielded no matches. Table 4 shows the Group 2 reports of hours saved for 13 batches. The number of saved hours is commensurate to the allotted time given to these experts (i.e., much more variable timespan). Thus, an average of over 38 hours was saved, minus the outlier of 640 hours. The comments accompanying this outlier expressed (a) an estimate of several weeks for the conventional task, and (b) a huge relief at receiving the aggregated search results.

Table 3

Time savings reported by Browsing Expert Group 1

Response Number	Hours Saved
1	4
2	2
3	8
4	16
5	1
6	0
7	10
8	6
9	0
10	4
11	0
12	1
13	2
14	4
15	10
16	8
17	16

Table 4

Time savings reported by Browsing Expert Group 2

Response Number	Hours Saved
1	112
2	40
3	56
4	16
5	16
6	16
7	32
8	16
9	40
10	40
11	640
12	40
13	40

Observations about the Paired Searchers

The paired searchers achieved a rate of query development that exceeded prior observed performance of a sole computational linguist searcher with machine translation. Pairing the computational linguist with a domain expert (with his/her language and domain competency) resulted in a considerable increase in the rate of profile query development, and in outstanding growth of the master profile query dictionary.

Our two domain experts performed well, but it was (initially) difficult for a browsing expert to adapt to the new domain expert role. For example, domain experts could estimate the precision of profile search results and determine cutoffs, but they had difficulty in selecting cutoff ranks due to the comprehensive scanning habits of their training as a browsing expert. If there were any relevant-looking documents below the 90% precision cutoff, the novice domain expert would be tempted to include these documents by increasing the reported cutoff, sacrificing precision in favor of recall (Manning, Raghavan, & Schütze, 2008, p. 142). We developed these guidelines for domain experts, to make them more successful:

1. Ensure that the review of profile search results is faster than the directed browsing activity
2. Understand that each profile represents a concept, and that search results above a cutoff should be relevant to that concept
3. Focus on the goal of aggregated search results, which is significantly different from directed browsing
4. Continuously think about additional words/terms that might contribute to an existing profile query, or to additional new queries
5. Continuously assess the value of each profile query and each word/term

Two key factors appeared to make exploratory search and profile query development most productive: (a) strong communication, and (b) initiative and risk taking. One example of strong communication was the domain expert reading aloud his/her review of foreign language profile search results, or vocalizing the matching content in full-document view. This allowed the pair to team up in determining an appropriate cutoff. The best rank selections resulted from collective team deliberation rather than from one searcher's decision. One example of initiative and risk taking was the domain expert suggesting a reformulation that was a significant departure from the theme of prior query iterations.

The domain experts varied significantly in the following skill areas: the speed of querying, the speed of document review, the thoroughness of document review, the tolerance for false positives above the cutoff, and inclination to experiment with the terms used for a profile.

Additionally, each domain expert learned, and adapted to, exploratory search at different speeds, which impacted the growth rate of the master profile query dictionary. The computational linguist adapted to the working style of each domain expert. This fostered teamwork among the paired searchers.

Observations about Semantics and Serendipitous Discovery

We observed that (a) attention to semantics and (b) serendipitous discovery were key to quality aggregated search results. The paired searchers ensured that the meaning of a profile (semantic class) was consistent with the meaning of the matching words. Frequently, the paired searchers recognized matching content that was present due to an irrelevant polysemic match. The word matched, but the meaning of the document content was outside the semantic class of the profile at hand; the query found unintended matching content. Occasionally, the match was correct, but the use of the matching word(s) was purely incidental, such as using the term merely as an example. Such matches were excluded from lists of profile search results by adjusting the cutoff, or by adding a query argument.

Extended exploratory search activity caused profiles to drift from the planned semantic class, or to become too specialized. Therefore, the profile name of each profile query was revisited periodically to validate that it corresponded to the semantic class of the query.

When paired searchers discovered content that responded to an information need outside the semantic class under consideration, the computational linguist noted the file name and content. Later, existing or new profiles were formulated to retrieve the serendipitously-discovered file.

Observations about the Balancing of Time Constraints and Exploratory Search

We found it challenging to make the paired searchers adhere to the two heuristics restricting labor and time: one hour per profile, and two days per batch. The more labor and time the searchers invested per profile, and completing a batch, the better the quality of the ensuing matrix and the greater its value. Furthermore, when the size of the master profile query dictionary exceeded 400 profiles, it became obvious that manually running that many profiles was consuming excessive time; easily including large quantities of profiles (e.g., hundreds) in aggregated search results is future work discussed below.

Observations about the Search Strategies and the Review of Search Results

The extensive Rhapsode operators allowed the paired searchers to learn the contents of each large batch, and to discover content that responded to the information needs, often serendipitously.

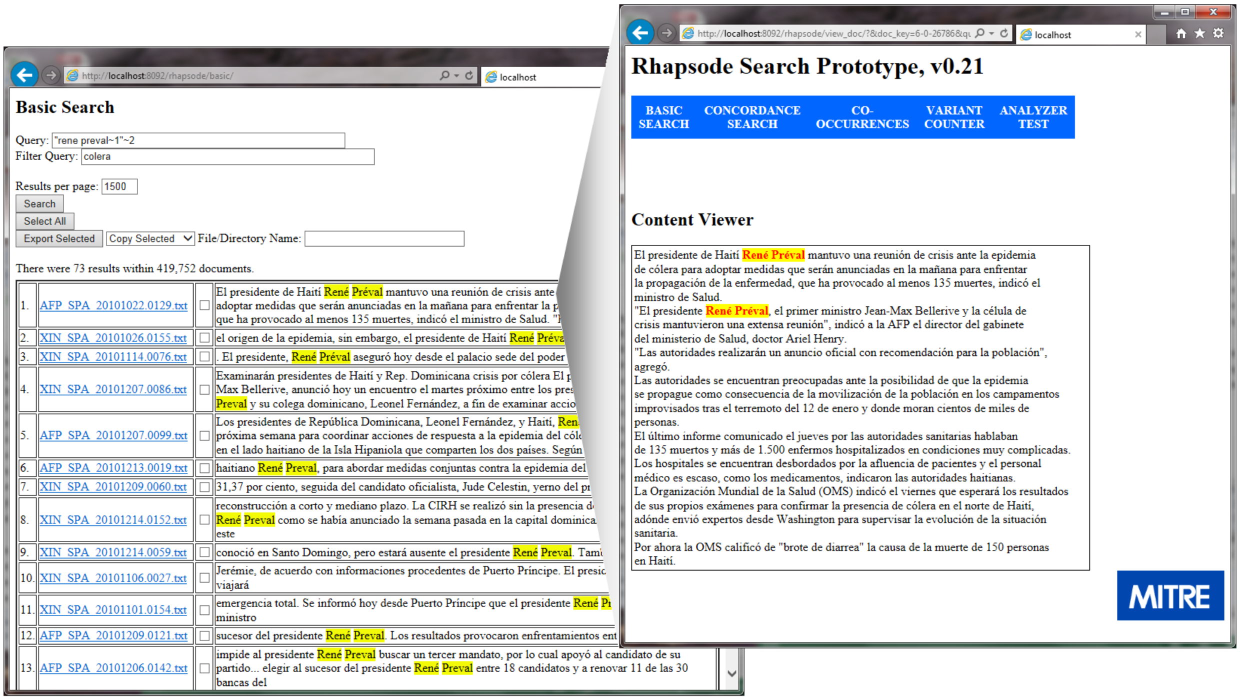
The tactics used by the paired searchers can be described in terms of Marchionini’s four general search strategies: building blocks (combining two or more individual concept queries), successive fractions (decomposing an information need into a set of queries), pearl growing (iterative search, where a query may be expanded on each iteration, using knowledge acquired on that iteration), and interactive scanning (iterative searching, incorporating review and/or comparison of retrieved documents, and other resources, toward a wide understanding of the topic) (Marchionini, 1997, pp. 77-80; Meadow and Cochrane, 1981, pp. 137-141; Harter, 1986, pp. 172-184; Hawkins and Wagers, 1982, pp. 12-13). Marchionini described interactive scanning as “guided discovery”.

The computational linguist ensured that the paired searchers employed a systematic searching style that is similar to interactive scanning. Building blocks, successive fractions, and pearl growing were incorporated as needed.

The paired searchers kept in mind the goal of aggregated search results when developing each individual profile query. By combining profiles, or subdividing profiles, they strove to obtain profile search results that were neither too long, nor too short, because they observed that attention to the length added to the overall quality of the aggregated search results.

A large part of developing aggregated search results was in reviewing profile search results. The length of the profile search results varied from query to query, depending on the number of files with matching content. Longer search results required more review time, driving the team to reduce the length of each (e.g., using more discriminating arguments), without losing important file matches. Paired searchers reduced search results that exceeded 30 files.

Figure 9 shows two Rhapsode Basic Search query fields: the query field and the filter query field. The latter was used extensively to reduce the number of search results.



Basic Search

Query: "rene preval~1"~2
Filter Query: colera

Results per page: 1500

File/Directory Name:

There were 73 results within 419,752 documents.

1.	AFP_SPA_20101022.0129.txt	□	El presidente de Haití René Préval mantuvo una reunión de crisis ante adoptar medidas que serán anunciadas en la mañana para enfrentar la p que ha provocado al menos 135 muertes, indicó el ministro de Salud. "I
2.	XIN_SPA_20101026.0155.txt	□	el origen de la epidemia, sin embargo, el presidente de Haití René Préva
3.	XIN_SPA_20101114.0076.txt	□	. El presidente, René Préval aseguró hoy desde el palacio sede del poder
4.	XIN_SPA_20101207.0086.txt	□	Examinarán presidentes de Haití y Rep. Dominicana crisis por cólera El p Max Bellerive, anunció hoy un encuentro el martes próximo entre los pres Preval y su colega dominicano, Leonel Fernández, a fin de examinar accio
5.	AFP_SPA_20101207.0099.txt	□	Los presidentes de República Dominicana, Leonel Fernández, y Haití, Ren próxima semana para coordinar acciones de respuesta a la epidemia del cóle en el lado haitiano de la Isla Hipaniola que comparten los dos países. Según
6.	AFP_SPA_20101213.0019.txt	□	haitiano René Preval, para abordar medidas conjuntas contra la epidemia del
7.	XIN_SPA_20101209.0060.txt	□	31,37 por ciento, seguida del candidato oficialista, Jude Celestin, yerno del pr
8.	XIN_SPA_20101214.0152.txt	□	reconstrucción a corto y mediano plazo. La CIRH se realizó sin la presencia d René Preval como se había anunciado la semana pasada en la capital dominica este
9.	XIN_SPA_20101214.0059.txt	□	conoció en Santo Domingo, pero estará ausente el presidente René Preval. Tami
10.	XIN_SPA_20101106.0027.txt	□	Jerémie, de acuerdo con informaciones procedentes de Puerto Príncipe. El presic viajará
11.	XIN_SPA_20101101.0154.txt	□	emergencia total. Se informó hoy desde Puerto Príncipe que el presidente René Pr ministro
12.	AFP_SPA_20101209.0121.txt	□	sucesor del presidente René Preval. Los resultados provocaron enfrentamientos ent
13.	AFP_SPA_20101206.0142.txt	□	impide al presidente René Preval buscar un tercer mandato, por lo cual apoyó al candidato de su partido... elegir al sucesor del presidente René Preval entre 18 candidatos y a renovar 11 de las 30 bancas del

Rhapsode Search Prototype, v0.21

Content Viewer

El presidente de Haití René Préval mantuvo una reunión de crisis ante la epidemia
de cólera para adoptar medidas que serán anunciadas en la mañana para enfrentar
la propagación de la enfermedad, que ha provocado al menos 135 muertes, indicó el
ministro de Salud.
"El presidente René Préval, el primer ministro Jean-Max Bellerive y la célula de
crisis mantuvieron una extensa reunión", indicó a la AFP el director del gabinete
del ministerio de Salud, doctor Ariel Henry.
"Las autoridades realizarán un anuncio oficial con recomendación para la población",
agregó.
Las autoridades se encuentran preocupadas ante la posibilidad de que la epidemia
se propague como consecuencia de la movilización de la población en los campamentos
improvisados tras el terremoto del 12 de enero y donde moran cientos de miles de
personas.
El último informe comunicado el jueves por las autoridades sanitarias hablaban
de 135 muertos y más de 1.500 enfermos hospitalizados en condiciones muy complicadas.
Los hospitales se encuentran desbordados por la afluencia de pacientes y el personal
médico es escaso, como los medicamentos, indicaron las autoridades haitianas.
La Organización Mundial de la Salud (OMS) indicó el viernes que esperará los resultados
de sus propios exámenes para confirmar la presencia de cólera en el norte de Haití,
adónde envió expertos desde Washington para supervisar la evolución de la situación
sanitaria.
Por ahora la OMS calificó de "brote de diarrea" la causa de la muerte de 150 personas
en Haití.

Figure 9. Rhapsode Basic Search and document viewer.

Observations about Aggregated Search Results

The aggregated search results presented an attractive advantage to Group 1 and a compelling advantage to Group 2 (English-only browsing experts). The aggregated search results clustered each batch of files by the information needs (i.e., by profile); files appeared with other files where their correlated profiles were similar. The resulting search matrix (or search grid) provided a visualization of all clusters. Because this was an Excel spreadsheet, users could modify this matrix visualization of clusters by sorting on selected combinations of profiles.

The aggregated search results transformed each batch of files (unstructured content) into structured content, or *profile metadata*, which provided a quick summary of each file. Profile metadata could be indexed and searched, used for data mining, or be piped to a visualization tool for browsing. Although the profile queries behind profile metadata vary from batch to batch, *controlled profile queries*, kept constant across all batches, would allow cross-batch correlation.

The nature of the aggregated search results appeared to present the opportunity for a user to perform a building blocks or successive fractions search, within Excel, independent of any search software. A building blocks search would involve systematically combining columns, each representing an individual set of profile search results, followed by re-sorting. Whereas, a successive fractions search would involve systematically removing columns.

We learned that distributing the static/immutable subsets of the master profile query dictionary, and the Tika-text-extracted renditions of the original files, was important to augment the users' understanding of the aggregated search results.

Discussion

The direct feedback from browsing experts via questionnaires provided an important source of information to guide and refine the hybrid strategy. It also presented a challenge in that Group 1 was evaluating a process and a knowledge product that was not yet part of the conventional procedure. These factors lowered Group 1's ratings. There was an observable rating increase halfway into the pilot implementation. Perhaps Group 1 became more familiar with the hybrid strategy and the aggregated search results, or the paired searchers refined their skillset by this point.

High ratings from Group 2 demonstrated that the aggregated search results added value to browsing experts who spoke only English. These artifacts helped them overcome the language barrier and gave them visibility into content that was otherwise obscure to them.

The master profile query dictionary is a core enduring asset, with a recurring return on investment as it is used, reused, and translated to other languages. For a new batch in the same domain, the dictionary can be re-used; for a new language, it can be translated. This dictionary can be used in two ways: fully-automated, or human-in-the-loop. The former uses the dictionary as is for batch processing. The latter allows the paired searchers to refine and/or add profiles to address the effects of lexical changes over time, and semantic drift.

The following rough calculation illustrates that the 640 hours saved outlier is understandable. Pirolli and Card (1999, p.648, 650) described two information overload scenarios: (a) an analyst faced with reading 34,000 pages per month from trade magazines, and (b) a couple of students faced with reading 300 papers retrieved from a digital library. The participants systematically filtered and reduced the information space, sacrificing coverage, so that they could accomplish their task, reading at a rate of 200 wpm. "Typically, analysts cannot explore all of the space and must forgo coverage in order to actually enrich and exploit the information" (Pirolli, 2007, p.190). In contrast, consider the browsing expert's task. For simplicity, assume (a) that human fatigue is not a factor, (b) that each file is comparable to a 900-word magazine page (as in Pirolli and Card), (c) that the browsing speed is 600 wpm (Carver, 1992). Thus, processing 100,000 files would take 2500 hours (62.5 weeks at 40 hours per week). Restricting this to two work days allows a review of 640 files (under 1% of the 100,000 files).

Cleverley, Burnett, and Muir (2017) concluded that organizations faced with information overload "may not 'know' they 'don't know'" that they are missing high value information (pp. 1, 93). Organizations need to assess the costs of partial and inaccurate coverage of the information space (an information gap that justifies the full hybrid strategy). A cost-benefit analysis may clarify the balance of labor between exploratory search and browsing. For the hybrid strategy to become fully operational, there is a cost to adding paired searchers and it is justified by the advantage of the strategy.

Amershi and Morris (2009) performed a study of co-located collaborative Web search and reported group frustrations when a single group member controlled the search application. In our study, paired searchers were unaffected by control; they were very collegial, relied on each other's unique skill sets, worked under a common goal with clearly delineated roles, and were highly motivated by professional rewards. Furthermore, they took ownership of their information seeking duties, providing a quality *service* to browsing experts.

Shah, Pickens, and Golovchinsky (2010) envisioned information seeking between two collaborating searchers, each with a complementary role. Their algorithmic mediation distributed search results tailored to each role, for a single iteration of a search session (no participants). In contrast, we observed professional participants in actual asymmetric roles, assigned by the skills they brought. Shah et al.'s setup involved two terminals/PCs, whereas our paired searchers used a single PC.

White (2016, p. 128) stated that in exploratory search, searchers are engaged in sensemaking; therefore, our study may appear to involve a sensemaking handoff from paired searchers to browsing experts. Sharma and Furnas (2009) compared the performance of synchronous collaborative searching with that of asynchronous collaborative searching (using the handoff of artifacts), and found the latter to be almost as effective as the former. In our study, the handoff of artifacts involved a high degree of common ground, awareness, and co-location. Browsing experts were free to contact the paired searchers for clarification, but chose not to. Paul and Morris (2009) reported successful asynchronous collaborative Web search handoffs using strategic awareness tools (query and chat logs). Our browsing experts were fully aware of paired searchers' goals via the initial discussion of information needs and the provided artifacts. An open research question is: would browsing experts use additional awareness aids, or ignore them due to the significant time pressure? For Sharma and Furnas (2009), and Paul and Morris (2009), initial searcher(s) serially handed off an unfinished knowledge product to a final searcher responsible to complete the task. In contrast, our paired searchers handed off finished knowledge products specifically tailored to aid a browsing expert. In providing the aggregated search results to the browsing experts, the hybrid strategy eased their sensemaking load.

Limitations of the Study

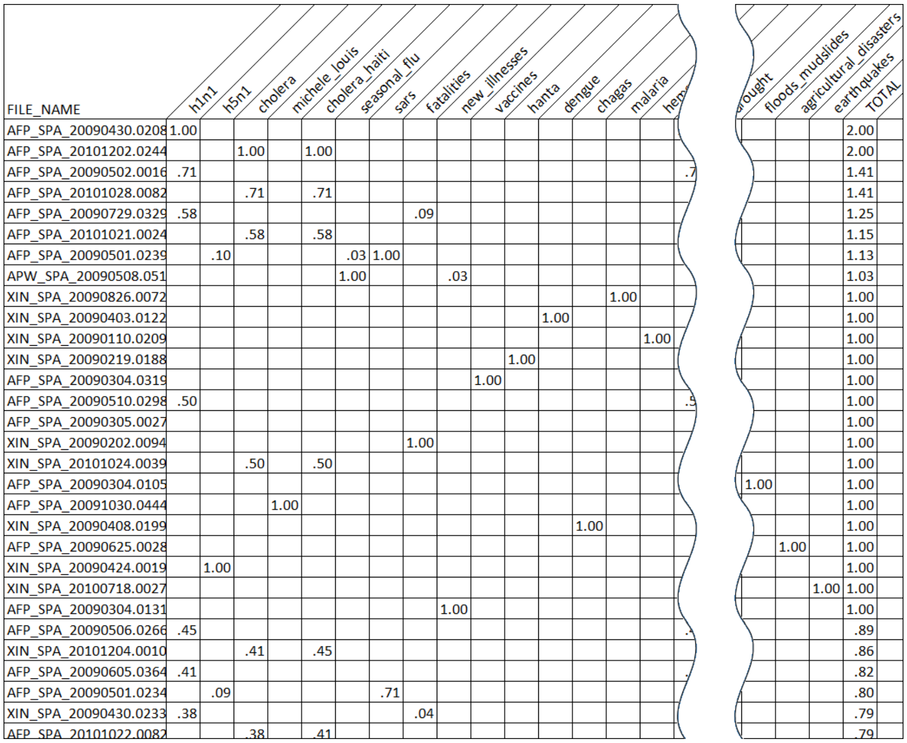
We acknowledge some limitations of the present study. First, we were observers in a real-world setting and lacked controls on the environment, which would be regulated in a laboratory setting (e.g., ground truth reference data, IR measures, time measures, etc.); therefore, our results are primarily qualitative. Second, the minimally-intrusive survey responses provided limited quantitative data, without a baseline, resulting in few quantitative findings and conclusions. Third, a small number of Likert items and a single time estimate provided limited insight into the overall gains, requiring a more in-depth and controlled evaluation against a different model; furthermore, the survey response rate was low. Finally, self-reports have questionable reliability; at the same time, these estimates were provided by highly experienced professionals. More research is needed to fully understand the benefits of the strategy for collaborative exploratory search.

However, the limitations of this study were compensated by strong qualitative observations; it was performed with professional browsing experts in a real operational setting with real challenges. The survey captured the sentiments of highly-motivated workers using a new knowledge product to facilitate asynchronous work. Additionally, the observed work culture was characterized by the transparent voicing of opinions, which adds to our confidence in the questionnaire responses.

Future Work: Rhapsode

The hybrid strategy can directly benefit from advancements to the search system. A subsequent version of Rhapsode now allows the user to manage and store profile queries inside the tool itself (including specification of cutoffs), and allows him/her to automatically generate aggregated search results. Because the aggregated search results Excel spreadsheet is accompanied by HTML renditions of each file with hit-highlighting, review of results only requires Excel and a web browser (independent of Rhapsode). The user clicks a cell of the aggregated search results and file content opens for review.

Rhapsode now preserves similarity score information in the aggregated search results. Rhapsode normalizes the real-valued Lucene similarity scores for each set of profile search results, and displays these in the aggregated search results (Figure 10), rather than using a “1” in each cell of the search matrix (Figure 5). We call these normalized scores *weighted relevance*.



FILE_NAME	h1n1	h5n1	cholera	michele_louis	cholera_haiti	seasonal_flu	sars	fatalities	new_illnesses	vaccines	hanta	dengue	chagas	malaria	hem		rought	floods_mudslides	agricultural_disasters	earthquakes	TOTAL	
AFP_SPA_20090430.0208	1.00																				2.00	
AFP_SPA_20101202.0244			1.00		1.00																2.00	
AFP_SPA_20090502.0016	.71															.7					1.41	
AFP_SPA_20101028.0082			.71		.71																1.41	
AFP_SPA_20090729.0329	.58							.09													1.25	
AFP_SPA_20101021.0024			.58		.58																1.15	
AFP_SPA_20090501.0239		.10				.03	1.00														1.13	
APW_SPA_20090508.051						1.00			.03												1.03	
XIN_SPA_20090826.0072														1.00							1.00	
XIN_SPA_20090403.0122												1.00									1.00	
XIN_SPA_20090110.0209															1.00						1.00	
XIN_SPA_20090219.0188											1.00										1.00	
AFP_SPA_20090304.0319										1.00											1.00	
AFP_SPA_20090510.0298	.50															.5					1.00	
AFP_SPA_20090305.0027																					1.00	
XIN_SPA_20090202.0094								1.00													1.00	
XIN_SPA_20101024.0039			.50		.50																1.00	
AFP_SPA_20090304.0105																		1.00			1.00	
AFP_SPA_20091030.0444				1.00																	1.00	
XIN_SPA_20090408.0199													1.00								1.00	
AFP_SPA_20090625.0028																			1.00		1.00	
XIN_SPA_20090424.0019		1.00																			1.00	
XIN_SPA_20100718.0027																				1.00	1.00	
AFP_SPA_20090304.0131									1.00												1.00	
AFP_SPA_20090506.0266	.45															.					.89	
XIN_SPA_20101204.0010			.41		.45																.86	
AFP_SPA_20090605.0364	.41															.					.82	
AFP_SPA_20090501.0234		.09					.71														.80	
XIN_SPA_20090430.0233	.38							.04													.79	
AFP_SPA_20101022.0082			.38		.41																.79	

Figure 10. Aggregated search results with *weighted relevance* (Example).

We intend to investigate the use of weighted relevance values for mass, automated generation of aggregated search results that obviates human estimation of cutoff ranks. If the full set of profile search results (each with no cutoff) is automatically consolidated into a set of aggregated search results, and if these results are sorted by weighted relevance values, a browsing expert could work down the matrix and abandon it when he/she encountered false positives indicated by low weighted relevance values.

We have also incorporated immediate reporting of profile statistics into Rhapsode, and intend to explore its benefits. On loading a batch and a master profile query dictionary, Rhapsode responds with a table, where each row indicates (a) the profile name, and (b) the count of search results corresponding to that profile query. This indicates entry points for exploratory search activity.

We plan to enhance Rhapsode with more NLP methods and human language technologies (HLT), many of which are underused because they have found little practical use on their own. We believe that search is a platform for increasing their use. Herceg and Ball (2010) discuss relevant use cases. We are exploring the use of machine learning models, based on a batch or massive corpora, to suggest term and phrase synonyms.

Future Work: The Hybrid Strategy

We intend to test the hypothesis that the hybrid strategy allows a search team to quickly adapt to shifts in the primary language of the incoming batches. This presupposes that the information needs and domain remain stable, and that the master profile query dictionary has been translated into the new language.

We are investigating different models for scaling up the strategy in an economical way. For example, we plan to observe a single computational linguist rotating, in a round robin manner, among two or more domain experts.

We seek (a) to study the filtering and triage of incoming batches over long periods, with the fully-automated application of a large master profile query dictionary, (b) to learn the characteristics of profiles that make them useful across batches, and (c) to learn the characteristics that cause profiles to need human review and refinement, or reformulation. Full automation rapidly transforms each batch of unstructured data into structured data (aggregated search results with weighted relevance). We intend to study (a) the usefulness of accumulated collections of aggregated search results for data mining, and (b) how novice users (i.e., those not skilled in data mining) would constructively use very large sets of automatically-generated aggregated search results, or accumulated collections of aggregated search results.

Our research persuades us that we should pursue information filtering of image, audio, and video data in the same manner as the textual media processing described in this paper. This presents an interesting research question: if the hybrid strategy can be used for all data types, then it provides a common framework for the filtering and triage of large-scale data in general. Granted, each data type necessitates media-specific preprocessing and media-specific search technology. Preliminary tests of the hybrid strategy on audio using automatic speech recognition and spoken content retrieval showed promising results.

Conclusion

Information seekers in many industries need mature data reduction mechanisms that filter and triage the daily flood of large, heterogeneous data sets. In this paper, we demonstrated the *hybrid strategy,* which allows comprehensive, multifaceted, detailed discovery in large unstructured data sets, in any language. We demonstrated that paired searchers with asymmetric skills and roles overcome the complexity of multilingual search and improve subsequent directed browsing. We introduced *aggregated search results* as a dividing layer that asynchronously and serially connects *collaborative exploratory search* and *directed browsing*. We observed a large positive effect from this serialization. Browsing experts provided Likert item ratings that reflected efficiency and effectiveness gains in terms of utility, speed, and quality. As such, 80% of Group 1 ratings and 100% of Group 2 ratings were positive. Also, we described the details of each component so that the strategy is fully reproducible.

The contribution is three-fold. First, in contrast to Web search, the hybrid strategy fits high performing organizations who apply human and machine resources to effect strategic information outcomes, despite an environment of exponential data growth (information overload) and severe time constraints. By serialization, the strategy overcomes the low quality of applying exploratory search alone (Cleverly, et al., 2017, p.92) and the low quality of applying directed browsing alone (see rough calculation above), thus increasing the yield of relevant files.

Second, the strategy provides significant data reduction. This reduction is facilitated by aggregated search results, which provides pointers to files containing relevant information by sorting the profile frequency counts (prioritization). This has a positive impact on the browsing expert who, in the restricted allotted time, can now focus on the most relevant information.

Third, the strategy divides information seeking into simplified subtasks corresponding to three distinct roles: computational linguist, domain expert, and browsing expert. This is important for business process optimization and for scaling to large information loads. Moreover, this division of labor offered an attractive advantage to browsing experts in general, and an especially compelling new solution for browsing experts unskilled in the language of the incoming data.

Acknowledgments

We wish to thank Stacey M. Bailey, Vanesa Jurica, Lori M. Lataillade, Jeffrey M. Siems, John S. White, George V. Wilson, and the anonymous JASIST reviewers. This work was supported by Contract W15P7T-13-C-A802, Department of the Army–Aberdeen Proving Ground; Approved for Public Release, Case#16-1413. The views, opinions, and/or findings contained in this report are those of The MITRE Corporation and should not be construed as an official government position, policy, or decision, unless designated by other documentation.

References

Allison, T.B. & Herceg, P.M. (2015). Methods for evaluating text extraction toolkits: An exploratory investigation. Technical Report MTR140443R1. January 22, 2015. McLean, VA: The MITRE Corporation.

Amershi, S., & Morris, M.R. (2009). Co-located collaborative web search: understanding status quo practices. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems* (pp.3637-3642). ACM.

Badi, R., Bae, S., Moore, J.M., Meintanis, K., Zacchi, A., Hsieh, H., Shipman, F.M. & Marshall, C.C. (2006, January). Recognizing user interest and document value from reading and organizing activities in document triage. In *Proceedings of the 11th international conference on intelligent user interfaces* (pp.218-225).

Bae, S., Badi, R., Meintanis, K., Moore, J.M., Zacchi, A., Hsieh, H., Marshall, C.C., & Shipman, F.M. (2005). Effects of display configurations on document triage. In *Human-Computer Interaction–INTERACT2005: IFIP TC 13 International Conference, Rome, Italy, September 12-16, 2005, Proceedings* (Vol. 3585, p.130). Springer.

Bae, S., Kim, D., Meintanis, K., Moore, J.M., Zacchi, A., Shipman, F., Hseih, H., & Marshall, C.C. (2010, June). Supporting document triage via annotation-based multi-application visualizations. In *Proceedings of the 10th annual joint conference on Digital libraries* (pp.177-186). ACM.

Belkin, N.J., & Croft, W.B. (1992). Information filtering and information retrieval: Two sides of the same coin? *Communications of the ACM*, 35(12),29-38.

Blair, D.C., & Maron, M.E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM*, 28(3), 289-299.

Buchanan, G., & Loizides, F. (2007). Investigating document triage on paper and electronic media. In *ECDL'07 Proceedings of the 11th European conference on Research and Advanced Technology for Digital Libraries*, (pp.416-427). Springer.

Buchanan, G., & Owen, T. (2008, October). Improving skim reading for document triage. In *Proceedings of the second international symposium on information interaction in context* (pp.83-88). ACM.

Büttcher, S., Clarke, C.L., and Cormack, G.V. (2010). *Information Retrieval: Implementing and Evaluating Search Engines*. Cambridge, MA: The MIT Press.

Carver, R.P. (1992). Reading rate: Theory, research, and practical implications. *Journal of Reading*, 36(2),84-95.

Case, D.O. (2012). *Looking for information: A survey of research on information seeking, needs and behavior, Third Edition*. Emerald Group Publishing.

Cleverley, P., Burnett, S., & Muir, L. (2017). Exploratory information searching in the enterprise: A study of user satisfaction and task performance. In *Journal of the Association for Information Science and Technology* 68(1). John Wiley & Sons.

Harter, S.P. (1986). *Online information retrieval: concepts, principles, and techniques*. Orlando, FL: Academic Press, Inc.

Hawkins, D.T., & Wagers, R. (1982). Online bibliographic search strategy development. *Online*, 6(3),12-19.

Herceg, P.M., & Ball, C.N. (2010). Reliable electronic text: The elusive prerequisite for a host of human language technologies. Technical Report MTR100302. September 30, 2010. McLean, VA: The MITRE Corporation.

Kelly, R., & Payne, S.J. (2013). Division of labour in collaborative information seeking: Current approaches and future directions. In *Proceedings of the International Workshop on Collaborative Information Seeking.*

Loizides, F., & Buchanan, G. (2009). An empirical study of user navigation during document triage. In *Research and Advanced Technology for Digital Libraries* (pp.138-149). Springer.

Macskassy, S.A., & Provost, F. (2001). Intelligent information triage. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* (pp.318-326). ACM.

Maiya, A.S., Thompson, J.P., Loaiza-Lemos, F., & Rolfe, R.M. (2013). Exploratory analysis of highly heterogeneous document collections. In *Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining* (pp.1375-1383).

Manning, C.D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge: Cambridge University Press.

Marchionini, G. (1997). *Information seeking in electronic environments*. New York: Cambridge University Press.

Marchionini, G. (2006). Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4),41-46.

Marshall, C.C., & Shipman, F.M. (1997). Spatial hypertext and the practice of information triage. In *Proceedings of the eighth ACM conference on Hypertext* (pp.124-133). ACM.

Meadow, C.T., & Cochrane, P.A. (1981). *Basics of online searching*. New York: John Wiley & Sons, Inc.

Mendonça, A., Daniel, J., Graff, D., & DiPersio, D. (2011). Spanish gigaword third edition LDC2011T12. Philadelphia: The Linguistic Data Consortium.

Morris, D., Morris, M.R., & Venolia, G. (2008). SearchBar: a search-centric web history for task resumption and information re-finding. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp.1207-1216). ACM.

Paul, S. A., & Morris, M.R. (2009). CoSense: Enhancing sensemaking for collaborative web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp.1771-1780). ACM.

Pickens, J., & Golovchinsky, G. (2007). Collaborative exploratory search. In *Proc. 2007 HCIR Workshop* (pp.21-22).

Pirolli, P. (2007). *Information foraging theory: Adaptive interaction with information*. Oxford University Press.

Pirolli, P., & Card, S. (1999). Information foraging. *Psychological review*, 106(4),643-675.

Pirolli, P., & Card, S. (2005). The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis* (Vol.5, pp.2-4).

Rogers, E.M. (2003). *Diffusion of innovations*. New York: Free Press.

Shah, C., Pickens, J., & Golovchinsky, G. (2010). Role-based results redistribution for collaborative information retrieval. *Information processing & management,* 46(6),773-781.

Sharma, N., & Furnas, G. (2009). Artifact usefulness and usage in sensemaking handoffs. *Proceedings of the American Society for Information Science and Technology,* 46(1),1-19.

Venolia, G. (2008). Backstory: A search tool for software developers supporting scalable sensemaking. Microsoft Research Technical Report MSR-TR-2008-13.

Voorhees, E.M., & Harman, D.K. (Eds.). (2005). *TREC: Experiment and evaluation in information retrieval*. Cambridge, MA: MIT press.

White, R.W., & Roth, R.A. (2009). Exploratory search: Beyond the query-response paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services, 1*(1),1-98. Morgan & Claypool Publishers.

White, R.W. (2016). *Interactions with search systems*. Cambridge University Press.