Managing Déjà Vu: Collection Building for the Identification of Nonidentical Duplicate Documents

Jack G. Conrad
Research & Development, Thomson Legal & Regulatory, 610 Opperman Drive, St. Paul, MN 55123.
E-mail: Jack.G.Conrad@Thomson.com

Cindy P. Schriber
Business & Information News, Thomson–West, 610 Opperman Drive, St. Paul, MN 55123.
E-mail: Cindy.Schriber@Thomson.com

As online document collections continue to expand, both on the Web and in proprietary environments, the need for duplicate detection becomes more critical. Few users wish to retrieve search results consisting of sets of duplicate documents, whether identical duplicates or close variants. The goal of this work is to facilitate (a) investigations into the phenomenon of near duplicates and (b) algorithmic approaches to minimizing its deleterious effect on search results. Harnessing the expertise of both client-users and professional searchers, we establish principled methods to generate a test collection for identifying and handling nonidentical duplicate documents. We subsequently examine a flexible method of characterizing and comparing documents to permit the identification of near duplicates. This method has produced promising results following an extensive evaluation using a production-based test collection created by domain experts.

Introduction

Both on the World Wide Web and in privately administered data environments, it is currently possible to have tens of millions of documents or more indexed as part of the same collection.[1] News databases are particularly challenging in that, thanks to news-wire pieces that are published by different newspapers, these databases may contain dozens of copies of the same article. A number of other domains also produce similarly large collections where the content of one document may be completely duplicated in another (Miller, 1990). These domains include business and finance, science and technology, medicine and bioinformatics, and intellectual property (Tenopir & Cahn, 1994). Within very large data environments like Westlaw and Dialog, in addition to the Web, environments which by themselves possess tens or hundreds of terabytes of data, the identification of duplicate documents is an important factor when developing practical and robust data-delivery platforms.

The problem of nonidentical duplicate document detection is just one example of the challenges faced by industrial Research & Development groups like the one at Thomson Legal & Regulatory (TLR). Such challenges often stand in contrast to those addressed by the purer research performed in more traditional or academic settings. Moreover, noteworthy differences between "industrial" and "pure" research would likely include the following distinctions: (a) Much pure research is too general for specific problems, (b) much pure research is little concerned with issues of scale, and (c) much pure research is driven by ideas rather than needs. The TLR R&D group is responsible for delivering production strength solutions to address real problems in customer-facing applications (Jackson & Moulinier, 2002). By focusing on the specific problem documented in this article and its large domain-driven application, and by putting users and their real-world needs in our line of sight, we avoid many of the limitations of pure research and adopt a vision which is clearly user centric.

In accepting the Gerald Salton Award at the 1997 SIGIR conference,[2] Tekfo Saracevic (1997) emphasized the importance of such a fundamental priority:

> The success or failure of any interactive system and technology is contingent on the extent to which user issues, the human factors, are addressed right from the beginning to the

[1] In this article, we use "collection" to refer to a database of textual documents.

[2] SIGIR is an Association of Computing Machinery (ACM) affiliated special interest research group. SIGIR stands for Special Interest Group in Information Retrieval (see www.acm.org/sigir/). The Gerald Salton Award is given for excellence in information retrieval-related research.

Received October 15, 2004; revised April 27, 2005; accepted May 9, 2005

© 2006 Wiley Periodicals, Inc. • Published online 00 XXXXXX 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20363

> very end, right from the theory, conceptualization, and design process on to development, evaluation, and to provision of services. (p. 26)

One of the strengths of the work reported on in this article is that it extensively enlists actual human practitioners into its research framework, from the beginning—the problem definition stage (via user representatives)—to the end—the query-result delivery stage (via professional assessors). In addition, procedures are validated for completeness and reliability through analyses of assessor agreement (using metrics of consistency, such as the kappa statistic), error rates, and significance. By employing such measures, the authors attempt to better characterize the degree of duplication in large, operational textual collections, and in so doing, to determine the scope of the underlying problem.

This work makes one fundamental contribution. It creates a deduping test collection by harnessing (a) real user queries, (b) a massive collection from an operational setting, and (c) professional assessors possessing substantial knowledge of the domain and its clients.

In addition, this work expands the discussion of online (real-time) deduping in Cooper, Coden and Brown (2002). Other recent work has often been syntax rather than lexical based, Web based (focusing on issues such as URL replication and instability), and conducted offline (e.g., examining large numbers of permutations before constructing a feature set). Previous research is thus substantially different than our current efforts which target a dynamic production environment.

The remainder of this article is organized as follows: We first review related work in duplicate document detection. Then we present the methodology used to assemble our duplicate document detection collection. Next, we describe an initial deduping algorithm for nonidentical duplicates and the preliminary trials to evaluate it, and then discuss performance issues associated with the algorithm. Following that discussion, we draw our conclusions and address future work. Finally, in the Appendices, we include samples of valuable user input and illustrations of near duplicates.

Previous Work

General Overview

Loosely related reports have proposed and theorized ideal test collections (Jones & van Rijsbergen, 1975) while others have analyzed existing collections to characterize a certain phenomenon such as interdocument duplication (Sanderson, 1997). More recently, efforts have been made to construct utilitarian, domain-specific collections (Hersh, Buckley, Leone, & Hickman, 1994; Shaw, Wood, Wood, & Tibbo, 1991), global, preclassified news collections (Rose, Stevenson, & Whitehead, 2002), corpora that facilitate specific tasks such as multilingual information retrieval (Sheridan, Ballerini, & Schäuble, 1996), summarization (Marcu, 1999), or filtering (Soboroff & Robertson, 2003) for Text REtrieval Conference (TREC) (Voorhees & Harman, 2000), and large information retrieval (IR) test collections (Cormack, Palmer, & Clark, 1998). This work appears to be the first to focus on a means of testing "fuzzy" (i.e., nonidentical) duplicate documents while making considerable efforts to satisfy expressed user preferences, thus bridging the gap between fully automated and user-centric identification (Saracevic, 1997).

Earlier Studies

Some of the first duplicate document detection studies addressed problems such as plagiarism, intellectual property violations, and partial replications within file systems (Brin, Davis, & García-Molina, 1995; Heintze, 1996; Manber, 1994). In many of these instances, researchers either owned or constructed their own datasets for the purposes of testing.

Concerning publicly available collections, in a published technical report, Sanderson (1997) described a set of tests he developed for the identification and potential removal of duplicate documents present in the Reuters test collection of over 22,000 news articles.[3] He performed a series of three tests to determine:

1. documents that are highly similar, but reported as separate events;
2. documents that are very similar, where one is a longer version of the other; and
3. documents that are exact duplicates of each other.

Candidate documents were found by submitting a document as a distinct query and examining the results. Documents were considered exact duplicates if the first retrieved the second, and vice versa. To avoid retrieving too many similar documents about related but different events (e.g., financial transactions), a condition was established requiring candidate pairs to be published within a 48-hr window of each other.

For Test 1, of 33 candidates for similar article, different topics, 29 (88%) were not about different events. For Test 2, of 283 candidates, 139 (49%) turned out not to be longer versions of the other. For Test 3, of 322 candidates, 320 (99%) passed the exact duplicates test. By presenting these findings, Sanderson (1997) helped characterize the nature and scope of the duplication problem in collections of news documents. Note that a more comprehensive review of pre-Web duplicate document detection research can be found in Conrad, Guo, and Schriber (2003).

Recent Web-Based Approaches

Much of the dedicated duplicate document research performed in the last decade has focused on TREC data or ad hoc corpora constructed from informal collections of Web

[3] The results Sanderson (1997) reported apply to both the original Reuters collection of 22,173 documents and the newer Reuters collection of 21,578 documents: www.daviddlewis.com/resources/testcollections/reuters21578/.

pages (e.g., Chowdhury, Frieder, Grossman, & McCabe, 2002).

Broder, Glassman, Manasse, and Zweig (1997) authored a seminal work on clustering Web-based documents that are *syntactically* similar to address a number of issues involving document *resemblance* and *containment* (multiple hosts, versioning, different formats, dead links, slow access, subsumption, etc.).[4] They conducted tests on virtually all of the Web at that time (i.e., in 1996). The authors' technique has come to be known as *shingling* and is applied by representing a document as a series of simple numeric encodings representing an *n*-term window—or shingle—that is passed over a document to produce all possible shingles (e.g., for $n = 10$). They then use filtering techniques to retain every *m*th shingle (e.g., for $m = 25$), and, if necessary, select a subset of what remains by choosing the lowest *s* encoded shingles (e.g., for $s = 400$). This process produces a document "sketch." To further reduce the computational complexity involved in processing large collections such as the Web, the authors present a super-shingle technique that creates meta-sketches or sketches of sketches. Documents that have matching super-shingles thus have a sequence of sketches in common. Pairs of documents that have a high shingle-match coefficient (resemblance) are asserted to be close duplicates while pairs that have lower match coefficients are similar. The authors used a resemblance threshold of 50% in their tests. As subsequent comparative tests have shown, the more distilled or abstracted the representations, the greater the chance for error (Chowdhury et al., 2002; Cooper, Coden, & Brown, 2002).

This work was expanded upon by Fetterly, Manasse, and Najork (2003) in a subsequent set of Web-based experiments. They identified clusters of near-duplicate documents and tracked their stability over time. They relied upon "mega-shingles" to compute clusters of near-duplicate documents, defined as documents having at least two super-shingles in common (i.e., a common mega-shingle). These authors found that two documents that are 95% similar have an almost 90% chance of having a mega-shingle in common; yet, two documents that are 80% similar have only a 2.6% chance of having a mega-shingle in common. In contrast to Broder et al. (1997), Fetterly et al. (2003) determined that their mega-shingling near-duplicate identification approach (using a union-find data structure) had a run time that was almost linear in the number of documents, O(N).[5]

Another recent approach used by Schleimer, Wilkerson, and Aiken (2003) is known as "winnowing." Like shingling, it can be adapted to a subset of local document fingerprints created by hashing; unlike shingling, it is based on strings of characters rather than strings of tokens. As such, winnowing ignores knowledge of its particular application domain (news and finance) as well as standard English text (tokens and their rarity). In some respects, winnowing operates at a logical extreme of the fingerprint by hashing. It applies an appreciable amount of math to the digital signature problem, but without harnessing domain expertise, semantic knowledge, or even term distribution information. It may be effective for general Web-based information about which we may know little, but for specific domains for which we know quite a bit, it may work at a disadvantage.

Both of the aforementioned approaches rely on hash values for each document subsection, and both prune these hash values to reduce the number of comparisons that the algorithms must perform. The computational complexity and thus resultant efficiency of the schemes are therefore quite dependent on the manner and extent to which the pruning is performed. The more aggressive the pruning, the more efficient are the algorithms, at the cost of increasing the prospects for identifying false-positive duplicates.

Shivakumar and García-Molina (1998) described factors in identifying nearly identical documents on the Web for the benefit of Web crawlers and Web archivers. They consequently concentrated on computing pairwise document overlap among pages commonly found on the Web. Their workshop draft specified Web-based applications for the identification of near replicas: (a) *more efficient web-crawling*, focusing on speed and richer subsets rather than time-consumptive comprehensiveness; (2) *improved results ranking* (or re-ranking), inspecting the environments from which Web documents originate; and (3) *archiving Web documents*, enabling greater compression of shorter pages that replicate more complete document sets. The authors revealed that there is a much greater incidence of (a) server aliasing, (b) URL aliasing, and (c) replication of popular documents such as frequently asked questions and manuals than initially believed. Some of the resource-saving concepts they proposed have been harnessed by a number of Web search engines, including Google (Brin, Davis, & García-Molina, 1995).

In one of the most comprehensive works to date, Chowdhury et al. (2002) refined their collection statistic, idf-based deduping algorithm for efficiency and effectiveness on both Web-based and non-Web-based test collections. They also compared its performance to other state-of-the-art techniques such as shingling and super-shingling. The authors demonstrated that their approach, called I-Match, scales in terms of number of documents and works well for documents of diverse sizes. They claimed that in addition to improving accuracy over competing approaches such as shingling, it executes in one fifth of the time. The authors briefly described how the collection statistics for the algorithm can come from training collections in rapidly changing data environments.

In more recent work, Kołcz, Chowdhury, and Alspector (2004) offered an alternative to I-Match that relies upon a set of digital signatures for a document created from randomized subsets of the global lexicon. The motivation for this approach is to compensate for the case where the fraction of

[4] Intuitively, Broder et al. (1997) defined resemblance to mean being "roughly the same" and containment to mean being "roughly contained within."

[5] Broder et al.'s (1997) multistep process took 10 CPU days to treat 30 million documents while Fetterly et al. (2003) processed 150 million documents in a fraction of that time.

DOI: 10.1002/asi

terms participating in the I-Match signature (hash) relative to the terms in the lexicon used is small. The significance of the approach stems from the fact that I-Match may result in false-positive matches if a large document has a small term intersection with the lexicon used. The authors showed that this approach outperforms traditional I-Match, with an improvement in overall recall of 40 to 60%. An advantage of the scheme is its increased insensitivity to word permutations and its document-length independence; however, the authors did not quantify the additional cost associated with generating the multiple lexicons, creating the multiple $(K + 1)$ signatures, and comparing one $(K + 1)$ tuple with another.[6] Thus, the computational cost of this improved performance is implementation dependent. For noncritical applications such as that mentioned by the authors—reducing spam by a significant percentage in a large Internet Service Provider's e-mail system—the benefits of the technique may outweigh its costs and justify its deployment.

The recent Web-related research of Park, Pennock, Giles, and Krovetz (2002) relied heavily on the notion of lexical signatures, consisting of roughly five key identifying words in document, based either on their low *df* or high *tf* properties. What distinguishes this work is that its eight signature variations are designed and evaluated for their ability either to retrieve the associated document in question in the top ranks of a search result (*unique identification*) or to retrieve alternative relevant documents should the document be "lost" (e.g., due to a broken link) (*relevance properties*). They determined that hybrid signatures consisting of only a couple of low *df* terms plus several high *tf* or high *tf.idf* terms produce the most effective unique *and* relevant properties for Web-page signatures.

Cooper et al. (2002) discussed methods for finding identical as well as similar documents returned from Web-based and internal IBM enterprise searches. The techniques are based upon the creation of a digital signature composed of the sum of the hash codes of the "salient" terms found in a document. The document signatures are intended to provide a shorthand means of representing the top terms in documents to facilitate fast comparisons. Their tests generally rely upon a single query and may warrant more comprehensive evaluation. The authors described their approach as the "logical extreme of super-shingl[ing]," yet, characterizing a document by summing its Java hash codes for hundreds or more terms may raise questions about the principled, dependable nature of the technique.[7]

The significance of this overview is that there has not yet been established a standard IR test collection for duplicate document detection. As we approached the problem, this was our first necessary step since without a validated test collection, we could not have confidence in the approaches and performance measures that followed.

Methodology

Background

Initially, the TLR business unit responsible for Business & Information News (BIN) asked us for technologies to identify and treat duplicate documents. In response, Conrad et al. (2003) began characterizing the distribution of duplicate types across news collections and then proceeded to address the two largest categories of duplicates. At the time, the BIN repository consisted of roughly 60 million news documents.

Much effort has addressed issues surrounding relevance assessments in various contexts of IR over the years (Burgin, 1992; Cleverdon, 1970; Harter, 1996; Voorhees, 2000). At a certain level of abstraction, the task we eventually asked our assessors to perform is similar in function to those of standard relevance judgments. Given an initial target document (that may be viewed conceptually as a query), our assessors are asked to identify other documents in the same result set that satisfy the similarity metrics (i.e., are "highly relevant" to it) established by a group of our client representatives.

Problem Definition and Client Feedback

We began by conducting a two-phase feedback session with 25 members of our Users Group, also known as the Library Advisory Board, who represent a variety of our clients' enterprises and firms. Most of the group's formal training is from the discipline of Library Science. These Board members generally field the information needs of their enterprise's legal practitioners and engage in a variety of research projects for their staff. Each of these individuals serves the needs of between 25 and 250 users, so they effectively function as a group of "meta-level researchers." They tend to focus on both the analysis and the synthesis of legal research at widely varying degrees of granularity. As such, they are uniquely positioned to provide domain expertise in their focus areas and are an excellent group to consult. In all, 17 of the 25 meta-level researchers provided nontrivial replies to our suite of questions.

The objective of the feedback session was to generate descriptions, both qualitative and quantitative, of the nature of the most annoying duplicate documents. In this exercise, the first phase was designed to depict the scope of the problem (examining various illustrations of duplicate documents in result sets) while the second phase was formulated to achieve consensus among the participants while quantifying and validating that dimension of the problem space most warranting treatment. The exercise resulted in the following description: A nonidentical duplicate document pair consists of two documents that possess a terminology overlap of at least 80% and where one document does not vary in length from that of the other by more than ±20%. It was generally believed that to call documents with less than an 80%

[6] $(K + 1)$: 1 represents the original and complete I-Match signature and K represents the number of permutations of the original lexicon. Kołcz et al. (2004) experimented with K ranging from 1 to 10.

[7] The test to determine whether a technique is principled, in this case, depends upon whether it avoids leaving anything to chance or probabilistic uncertainty. In short, is the approach highly reliable?

terminology overlap duplicate would be problematic. Although such documents might adequately satisfy Broder et al.'s (1997) definition of *containment*, they could not reasonably satisfy a definition for *resemblance*, which was ultimately the focus of our attention (see Appendix A for illustrations of Advisory Board contributions).

Note that the 80% overlap condition does not imply that 80% of the shared text must be identical (see Appendix B for an illustration of how inexact the resemblance may be). In the example shown in Appendix B, this example, even though these articles differ substantially at the paragraph, sentence, phrase, and word levels (not to mention at the title level), the illustrations still satisfy the similarity conditions of our definition and would thus be judged as valid nonidentical duplicates.

These guidelines produced a working definition of "near-duplicate" documents with which we proceeded. Note that implicit in this definition is the fact that these relations are *not* transitive. That is to say, if Documents A and B are duplicates and B is 80% the length of Document A, and Documents B and C are duplicates and C is 80% the length of Document B, it does not follow that C also is a duplicate of A. In this instance, that is clearly not the case.

Collection Generation and Domain Expert Assessments

To test our approach, we selected a total of 100 real user information requests from our query logs. These logs originate in the production environment that is responsible for the largest percentage of duplicate documents: news, including financial. The queries were randomly selected with the exception that we required a results list of at least 20 documents.[8] A sample of these queries is shown in Table 1. The average query contained roughly five terms, excluding date and proximity operators. Each query was run using the Westlaw system, which provides both Boolean and natural language search capability, depending on the preference of the user (Turtle, 1994).

As the queries in Table 1 suggest, a sizable majority of our News database subscribers prefer to use Boolean rather than natural language queries, often due to the perceived control it offers users. The default results ranking for Boolean queries on Westnews is by date (i.e., reverse chronological order).

TABLE 1. Sample queries for nonidentical duplicates.

Type	News & Finance [DB: ALLNEWSPLUS]
NL	Pay reform for federal law enforcement officers
NL	"consumer fraud" "deceptive behavior" "unfair practice"
Bool	"medical malpractice" & "public citizen" (1/25/03+)
Bool	"natural gas" & storage & "all time low"
Bool	"Eastern Europ*" & support & US & Iraq (02/01/03)+
Bool	John/ 3 Ashcroft (1/25/03+)

After running these queries against the ALLNEWSPLUS database consisting of approximately 50 million comprehensive ALLNEWS articles and another 10 million frequently updated [NEWS]WIRES articles, we assembled the top 20 documents returned from each query. We had each set of 20 documents reviewed by two client research advisors to identify their duplicate sets.[9] This process produced standard training and test sets against which computational approaches would be compared.[10]

Details of document inspections. In this trial, we applied a definition of nonexact duplicate that was generated by a customer work group, the Advisory Board described previously. The resulting definition states that two documents are duplicates if they retain 80% of the same language and a shorter document is not less than 80% the length of a larger document.[11] To formally review the duplication status of the result sets, we assembled two teams of two client research advisors. The 100 queries were divided into two sets of 50 (Buckley & Voorhees, 2000), the first set to be used to train the algorithm and the second set to test it. The process by which the query results were judged was scheduled over 4 weeks time (as indicated in Table 2). During Week 1, results from the training queries were assessed for their duplication status. Each team reviewed the results from 25 queries, 5 queries per team per day. Although members of the same team reviewed the same results, they did so independently.

Date restrictors (excluding the current day) were added to the queries to help ensure that the assessors would be examining exactly the same documents in the same order. The assessors also had access to the term counts available in the core documents (which excluded publisher-added classification terms and other metadata as shown in Table 3). Week 2 served as an arbitration week. When members of the same team disagreed about a duplicate set, a member of the other

TABLE 2. Scheduling of assessments.

Assessor pair	Team A	Team B
Week 1	25 queries	25 queries
Week 2	Arbitrate for Team B	Arbitrate for Team A
Week 3	25 queries	25 queries
Week 4	Arbitrate for Team B	Arbitrate for Team A
Total	50 queries	50 queries
Combined	100 queries	

[8] Between 5 and 10% of our candidate queries were replaced because they had less than 20 documents returned, given an initial set of non-null results.

[9] Our client research advisors, who also happen to have law degrees, spend a significant portion of their work day fielding customer research questions over the telephone.

[10] "Training" is not used here in the Machine Learning sense involving automatic learning; rather, it signifies an initial round in which we were permitted to establish the algorithm's optimal parameter settings.

[11] (a) That is, 80% of the words in one document are contained in the other (in terms of overall *terminology* rather than individual term *frequency*). (b) For documents that do not meet a working threshold for similarity or *resemblance*, Broder et al. (1997) monitored a second, looser relationship described as *containment*.

TABLE 3. Metadata classifications and examples.

Tag	Topics	Sample classifications
N1	News Sector	Political; Crimes/Courts
R1	Region	Indonesia; SE Asia; Asia
G1	Government	Australia; Indonesia
M1	Market Sector	Consumer Electronics
I1	Industry	Aerospace/Defense
C1	Company (tkr)	UnitedHealth Group (UNH)
P1	Product	Automobiles; Heavy Equipment
T1	URL	*www.tlrg.com*

team would serve as an arbitrator or tie breaker. Weeks 3 and 4 were conducted in the same manner using the remaining 50 queries, thereby creating the test set. In this way, a virtual voting system was established. Every result set would thus be reviewed by a minimum of two assessors, and sometimes by three. This approach was intended to produce dependable judgments from the process.

To further help ensure judgment reliability and consistency, a six-page training document was prepared for the assessors. It included illustrations and detailed instructions. Examples of some of the instructions and heuristics provided in the training document are shown below.

- assessment criteria consist of the words and sentences in the articles only;
  - ○ format-related features can generally be ignored
  - ○ document artifacts discovered via term-browsing can generally be ignored
- document paragraphs need not possess completely identical content to satisfy the duplication identification criteria;
- titles and author names alone are not reliable indicators of duplication, but represent one piece of evidence among multiple pieces
- publication dates that differ by more than 3 months generally do not yield duplicate candidates
- an 80% similarity threshold implies that if one of the candidate documents is less than 80% the length of the other, the pair generally will not satisfy our definition of duplicates
  - ○ a threshold heuristic for duplicate identification would require, for instance, at least 4 of 5, 8 of 10, or 6 of 7½ paragraphs among two documents being nearly identical

In addition, a preliminary training exercise was developed for each team that included real user query-result sets and the opportunity for the participants to discuss their judgments as well as the granularity of their inspection. All four assessors participated in the same initial training session and were asked to apply their knowledge to the same pair of sample result sets. Training guidelines were amended as a result of the session to clarify the level of granularity of analysis necessary for the task. In general, the assessors found the training exercise quite instructive. As beneficial as this training round was, the assessors did not produce completely uniform judgments. For that reason, information and statistics about interassessor agreement can be found in the next section on interassessor agreement.

TABLE 4. Distribution of duplicates across queries.

Duplicate document detection	Training set	Test set
Total Queries	50	50
With Dup Sets	41	44
Without Dups Sets	9	6

Table 4 presents the number of queries that yielded duplicate sets in the trial. Some queries produced no duplicate sets–nine in the training set and six in the test set. These were retained for two main reasons: (a) They were produced by our random sampling and are therefore presumably representative, and (b) they can still be instructive in terms of false-positive sensitivity experiments since these queries should produce no duplicate sets. Of these sets with no duplicates, seven were encountered by Team A and eight by Team B. The interassessor agreement for no duplicates was high. Team A agreed on queries with nonduplicates in seven of seven instances while Team B agreed on these queries in 8 of 13 instances (before arbitration).

By contrast, Table 5 shows the distribution of duplicate sets by size. The queries for the test set produced slightly fewer duplicate sets, but also several larger duplicate sets consisting of four, five, or six documents. The assessors identified an average of 1.7 duplicate sets per query-result set. In total, 2,000 documents were examined. The mean length of the news documents returned during the two rounds was 796 terms (excluding publisher-supplied indexing terms).

Interassessor Agreement

When asked to verbally characterize the nature of the duplicate sets identified, in relation to exact duplicates, the assessors were in agreement that the sets they found spanned the identical—nonidentical duplicates spectrum.[12]

Of the 100 queries reviewed by a pair of assessors, 53 resulted in complete agreement between the assessors. Furthermore, Team A agreed on 72% of its duplicate sets while Team B agreed on 55% of its duplicate sets. Disagreements between assessors were resolved by means of a voting process, whereby one of the assessors from the opposite team served as an arbitrator and cast a third and tie-breaking judgment.

TABLE 5. Distribution of total resulting duplicate sets.

Duplicate set size	Training set (Frequency)	Test set (Frequency)
Pairs	68	64
Triplets	12	12
Quadruplets	8	2
Quintuplets	0	3
Sextuplets	0	1
Total	88	82

[12] In Conrad et al. (2003), the authors categorized and quantified into six classes the distribution of duplicates found in this collection.

TABLE 6. Kappa statistics for interassessor agreements for duplicate set identification [macro-averaged scores (micro-averaged scores in parentheses)].

Assessor pair	Team A	Team B
Week 1		
(First 25 Queries)	$\kappa = 0.8549$ (0.8738)	$\kappa = 0.7089$ (0.7144)
Week 3		
(Second 25 Queries)	$\kappa = 0.8312$ (0.8423)	$\kappa = 0.7484$ (0.6831)
Weeks 1 & 3		
(50 Queries)	$\kappa = 0.8443^*$ (0.8580)	$\kappa = 0.7304^+$ (0.6987)
Combined		
(100 Queries) (Teams A & B)	$\kappa = 0.7829$ (0.7784)	

We used the kappa statistic for nominally scaled data to compare our interassessor concordances over the 100 result sets (Siegel & Castellan, 1988). The kappa coefficient of agreement is the ratio of the proportion of times that the assessors agree (corrected for chance agreement) to the maximum proportion of times that the assessors could agree (corrected for chance agreement):

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{1}$$

where $P(A)$ is the proportion of times that the k assessors agree and $P(E)$ is the proportion of times that we would expect the k assessors to agree by chance. If there is complete agreement among the assessors, then $\kappa = 1$, whereas if there is no agreement (other than the agreement that would be expected by chance) among the assessors, then $\kappa = 0$. We used as our baseline set of candidate duplicates the set of all document pairs identified by at least one of our assessors. The results are presented in Table 6.[13]

Computational linguists have taken $\kappa = 0.8$ as the norm for significantly good agreement, although some have argued that there is insufficient evidence to choose 0.8 over, for instance, other values between 0.6 and 0.9 (Marcu, personal communication, April 24, 2002).

Given a result set of $n = 20$ documents, there are $n(n - 1)/2$ or 190 total comparisons required. We had two assessors make categorical judgments with respect to each of these candidate pairs: duplicate or nonduplicate. We computed the kappa statistic over the comparison space described.

Because the majority of document pairs are not duplicates, the possibility for chance agreement is high. But this marshals the strength of the kappa statistic—it corrects observed agreement with respect to chance agreement. Given the size of the space (190 pairwise comparisons), the resulting kappa values we obtained may be slightly inflated (given that the vast majority of the 190 comparisons are nonduplicates), but not significantly (Marcu, personal communication, December 16, 2003). Nonetheless, in three of four instances where a duplicate candidate was identified by one of the assessors, the two assessors agreed on its duplicate status.

After determining the value of the kappa statistic, κ, it is customary to determine whether the observed value is greater than the value which would be expected by chance. This can be done by calculating the value of the statistic z, where

$$z = \frac{\kappa}{\sqrt{\mathrm{var}(\kappa)}} \tag{2}$$

to test the hypothesis H_o: $\kappa = 0$ against the hypothesis H_1: $\kappa > 0$ (Carletta, 1996; Siegel & Castellan, 1988).

The value of κ for the combined query set yields $z = 1.965$ (Team A, Queries 1–50)* and $z = 1.842$ (Team B, Queries 51–100)$^+$. These values exceed the; $\alpha = 0.05$ significance level (where $z = 1.645$). Therefore, we may conclude that the assessors exhibit significant agreement on this categorization task. Note that these results were produced *before* we introduced the arbitration round, wherein a member of the alternate team resolved differences in judgments between the two original assessors. Given a third expert casting a "vote" on these differences, the final duplication judgments are arguably more reliable than those examined during the kappa analysis.

Overview of Initial Algorithm

The overview of initial algorithm and the following section about collection deployment and performance evaluation are included to demonstrate the utility of such a duplicate document detection collection when designing, developing, and testing algorithmic approaches to deduping.

Note that there have been efforts to completely automatically detect "redundancy" in result sets (Zhang, Callan, & Minka, 2002), but these appear to eliminate the role of the client and focus exclusively on mathematical models of content, even in highly dynamic retrieval environments. A comprehensive work that compares document similarity (or "identity") measures with fingerprint (or "hashing") approaches found that both can be used to effectively identify near duplicates, but also concluded that given the sensitivity of fingerprints, similarity measures are superior (Hoad & Zobel, 2002). Yet, such document-similarity measures require that every document be compared to every other document and are thus computationally prohibitive given its theoretical run time of O(N^2), where N is the size of the collection.[14]

To determine our ability to identify and characterize such nonidentical duplicate documents in our production

[13] For the macro-averaged scores, the kappa statistic is calculated using a single table for all the comparisons involved in the entire query set. The micro-averaged scores are calculated using a separate table for the comparisons from each query; these scores in turn are averaged together to derive the composite kappa score.

[14] Chowdhury (2004) mentioned that in reality, documents are only compared if they have overlapping terms, which reduces the runtime by a fraction that is difficult to predict.

DOI: 10.1002/asi

environment, using the input from our client base, we began investigating reliance upon an expanded multidimensional feature set or "digital signature." This feature set includes:

- time component (pub_date);
- magnitude component (doc_length);
- core content component (term_vector).

The role of the first two is to provide heuristics to reduce the need for more costly term comparisons (as per the last two bullets of the *Details of document inspections* section). They do not reduce the number of candidate pairs as much as reduce the search space for valid duplicate candidates. In addition to a publication date (e.g., days or weeks since January 1, 1950) and document length (excluding metadata), a document's term_vector is represented by its top *n* idf words, where *n* falls somewhere between 30 and 60 words. We determined empirically that 60 words would serve as an optimal default vector size because (a) it offers substantially finer granularity to the process, and (b) it does not exceed the short length limits of the vast majority of our shortest news documents. In a number of instances, there are not always many more than 60 terms to select for discrimination purposes.

The percent overlap between two documents' term_vectors served as our de facto similarity measure. In practice, once our heuristics are invoked and complete their reduction of eligible candidate pairs, the algorithm then uses as its matching criteria, a vector overlap of at least 80%.

Aside from core document content, metadata indicating region, news sector, market sector, industry, product type, and so on (shown in Table 3) are not used. We have determined that such categories tend to increase the number of false positives since related, but dissimilar, documents may possess similar metadata classification terms.

Note that even though these metadata classification indexes are not considered part of the core document, they were not suppressed from our assessors, though the assessors were generally discouraged from using them in their determination of duplication status because of the risk of false positive identification discussed earlier. Nonetheless, in the comprehensive collection that resulted, these fields are still viewed as intrinsic to the corpus, and therefore are retained.

Collection Deployment and Performance Evaluation

Test Corpus and Algorithm Assessment

Figures 1a and 1b show the performance of the algorithm outlined earlier relative to the standard established by the client research advisors, in terms of agreement (correct identification), false negatives (misses), and false positives (overgeneration). An idf table constructed from a training collection of over 2 million documents is used. A number of modifications were made to the algorithm during the training phase. Most notable is how it treats short documents (with fewer than 60 terms). A variety of options exist, including (a) comparing vectors of unequal length, (b) comparing only the rarest *n* terms, where *n* is the size of the shortest document's vector, and (c) padding the short document's term vector with entries not found in the table (in a manner that facilitates comparisons with similar documents). In the end, we found that amendments to the last approach yielded the best results. We nonetheless discovered that (atypical) documents of less than 20 terms yield a higher rate of false positives and thus are not recommended as reliable candidates for signature generation.

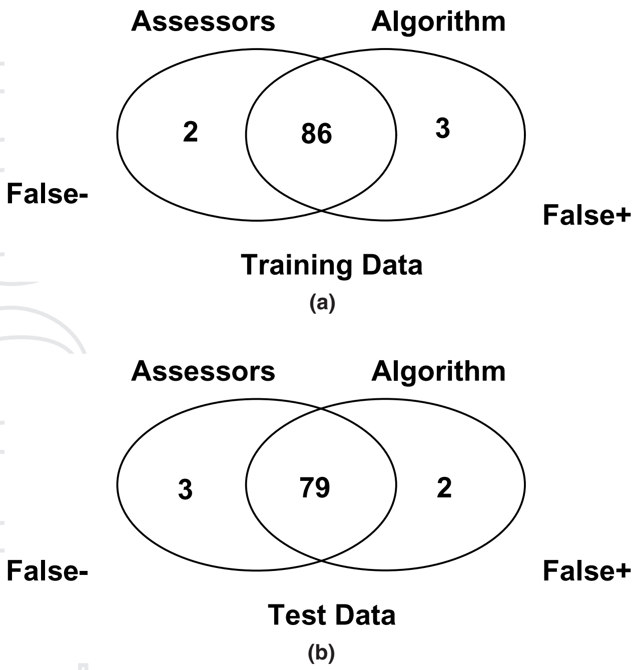

FIG. 1. (a) Duplicate sets identified in training round. (b) Duplicate sets identified in test round.

The algorithm recognized 98% of the duplicate sets identified by the assessors in the training round (86/88) with three false positives and 96% of the duplicate sets in the test round (79/82) with ultimately two false positives. Upon performing a failure analysis of our false positives, we were able to make three key observations. First, by tightening our date window for candidates from 12 weeks to 6 weeks, three of the test round's false positives can be eliminated with no impact on the other results. Second, three of the false positives are sets resulting from the same financial query about "cotton futures" and are either practically all numerical or represent a boiler plate text of nearly identical content with only two numbers changing from one day's document to the next.[15] The significance of these observations is that for those largely quantitative financial documents for which the algorithm was not designed, performance is spotty, and some user education may be helpful. Third, of the remaining

[15] In an IR context, the percentages presented correspond to recall. By contrast, 86 of 89 (96.5%) and 79 of 81 (97.5%) correspond to precision (excluding the three date and three numeric-resolved misses) (cf. Table 7).

TABLE 7. DDD Algorithm-assessor correspondence.

DDD Algorithm performance	Training set	Test set
Precision (%)	86 of 89 (96.5%)	79 of 81 (97.5%)
Recall (%)	86 of 88 (98.0%)	79 of 82 (96.0%)

false positives, the documents are often so close that the extent of their "erroneous" nature is debatable among the assessors (e.g., John Ashcroft's Justice Department budget presented before the House and before the Senate—same material, different audience, different title).

If we define *precision* as the percentage of duplicate documents identified by the algorithm that agree with the assessors and *recall* as the percentage of the total number of duplicate documents identified by the assessors also identified by the algorithm, then our results can be found in Table 7.

Comparative Evaluation

Any analysis like the one presented earlier would be incomplete without a discussion of comparative evaluation. The most useful comparison to examine is that of idf-based deduping techniques (discussed in the previous section) and well-known alternatives such as shingling (Broder et al., 1997), in terms of both timing and effectiveness. It is important to mention that when we incorporated features such as doc_ length and pub_date into the digital signature, they were selected in a manner to minimize impact on overall performance. That is, we selected a window (of time) and a range (of length) such that no duplicates would be lost by their introduction; they serve strictly to reduce the computational cost of comparisons. For this reason, the comprehensive trials conducted by Chowdhury et al. (2002) provided some useful insights. They examined how idf-based signature approaches to deduping perform relative to selective windowing techniques such as shingling. They determined that given identical data, an optimized idf fingerprint approach is nine times faster than shingling (six times faster than super-shingling) when run against the 2 GB NIST Web collection (on a Sun ES-450) (Hawking, Voorhees, Craswell, & Bailey, 2000).[16]

In terms of actual deduplication effectiveness, because shingling does not cover every portion of a textual document and is not sensitive to the rareness of participating terms, it consistently underidentified duplicates in a diverse duplicate set constructed from TREC's Los Angeles Times subcollection (Voorhees, 2000) (which consisted of 10 duplicate sets of 11 documents each). This outcome resulted as shingling produced more than the optimal number of duplicate sets when processing the automatically generated test collection. Although both approaches use principled techniques, a key distinction between them is that shingling relies upon undiscriminated strings of tokens (shingles) as its characteristic content (discussed earlier). By contrast, the idf-based algorithms distinguish between rarer, richer, content-bearing terms and those which are not. This appears to be one of the chief shortcomings of shingling and a strength of idf-based approaches.

Accuracy and Confidence Levels

Note that our evaluation of the algorithm's results on our test set provides only an approximation of its true accuracy. After all, we applied our algorithm to a combined sample set of 2,000 documents from a collection of over 50 million. A reasonable question is thus "how good of an approximation is this?" Stated differently, what is our confidence level that the performance measures on this set reflect true accuracy on the complete set? Mitchell (1997) addressed this problem in the context of Machine Learning. For a collection *C*, $error_C$ can be defined as the ratio of false positives and false negatives in the algorithm's results on *C*. Our evaluation test set of sample *S* produces $errors_S$. Mitchell assumes that the probability of having a specific ratio of errors (r) is approximated by a normally distributed random variable with a mean errors_S and standard deviation:

$$\sigma_{errors_S} = \frac{\sigma_r}{|S|} \approx \sqrt{\frac{errors_S(1 - errors_S)}{|S|}} \qquad (3)$$

where $|S|$ is the size of the sample. The true error can be viewed as drawing a bell curve that is centered on the observed error. So with probability *N*%, $error_C$ is within z_N *SD*s of $errors_S$, where z_N is the z value. In our case, there is a 95% chance that error_C is within 1.96 *SD*s of $errors_S$. For instance, for an observed error ratio of 0.5% (five errors among 1,000 documents), there is a 95% chance that the error on the full collection is within the range 0.50% ± 0.22%. For 10 errors among 2,000 documents, the interval is 0.50% ± 0.16%. This analysis likely warrants further investigation since as one moves beyond consideration of a result set consisting of 20 documents, the number of pairwise comparisons required per query increases exponentially. It would be instructive to determine whether this fanout has any appreciable impact on error rate. In subsequent tests on result sets consisting of approximately 1,000 documents, we found no deviation in performance.

Conclusions

The need for flexible approaches to duplicate document detection has become critical due to the explosive growth of globally distributed and accessible electronic data. This task can take a variety of forms, but can be fundamentally characterized as either identical or nonidentical duplicate detection, and also may depend on domain and method of search. We have explored a production-based replication problem that is prevalent in the news domain. We designed a methodology that invited meta-level users with backgrounds in

[16] This result is not wholly applicable to our environment since our initial algorithm performs actual term-based comparisons (on average, 15 per document-pair) when the heuristic filtering fails.

DOI: 10.1002/asi

library science to define the scope of the problem, and then commissioned two teams of professional searchers to use our working definition and additional principled methods to construct a new collection in which nonidentical duplicates are identified. We also have attempted to validate the decisions of our assessors using a kappa analysis. For nonidentical duplicate document detection, our applied test collection proved beneficial, and the subsequent trials suggest that a multidimensional feature set approach to characterizing and comparing documents can provide a strong indicator of the degree of duplication between two documents. The treatment of its multidimensional feature set frees it from reliance upon singular features and permits heuristics to save on more costly comparisons.

The ultimate goal of this work is to provide a useful, tractable, validated test collection to facilitate the design, development, and evaluation of algorithms for identifying and treating nonidentical duplicate documents in textual databases. Based on our experience with a trial deployment of this collection and the performance analyses that followed, we believe we have accomplished our prime objective.[17] In the absence of such a validated test collection, it would be difficult to establish confidence in the characteristics of a corpus and the resulting performance of any trial algorithms.

Future Work

We plan on further quantifying the error statistics in future work, in terms of the size of the result set. We will also address document deduplication strategies that extend beyond the results of individual queries and singular domains.

Acknowledgments

We appreciate the duplicate assessment efforts of Brian Craig, Carol Jo Lechtenberg, Michael Paulos, and Tom Perusse. We acknowledge the support of Marilee Winiarski, who invested in our nonidentical duplicate research. We also thank Bruce Getting, Jane Lund, Jie Lin, and Jeremy Leese for their handling of computability and real-time processing issues in the production environment. Finally, we thank the anonymous reviewers for their constructive comments and questions regarding this work.

References

Brin, S., Davis, J., & García-Molina, H. (1995). Copy detection mechanisms for digital documents. Proceedings of the Special Interest Group on Management of Data, San Francisco (pp. 398–409). ACM Press.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. Proceedings of the 7th International World Wide Web Conference, Brisbane, Australia (pp. 107–117). Essex, UK: Elsevier Science.

Broder, A., Glassman, S., Manasse, M., & Zweig, G. (1997). Syntactic clustering of the Web. Proceedings of the 6th International World Wide Web Conference, Santa Clara, CA (pp. 391–404). Essex, UK: Elsevier Science.

Buckley, C., & Voorhees, E.M. (2000). Evaluating evaluation measure stability. Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia (pp. 282–289). ACM Press.

Burgin, R. (1992). Variations in relevance judgments and the evaluation of retrieval performance. Information Processing and Management, 28(5), 619–627.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. Computational Linguistics, 22(2), 249–254.

Chowdhury, A. (2004). Duplicate data detection. Retrieved from http://ir.iit.edu/~abdur/ Research/Duplicate.html

Chowdhury, A., Frieder, O., Grossman, D., & McCabe, M.C. (2002). Collection statistics for fast duplicate document detection. ACM Transactions on Information Systems, 20(2), 171–191.

Cleverdon, C. (1970). The effect of variations in relevance assessments in comparative experimental tests of index languages [Tech. Report], Cranfield Library Report No. 3, Cranfield Institute of Technology, Cranfield, United Kingdom.

Conrad, J., Guo, X., & Schriber, C. (2003). Online duplicate document detection: Signature reliability in a dynamic retrieval environment. Proceedings of the 12th International Conference on Information and Knowledge Management, New Orleans, LA (pp. 443–452). ACM Press.

Cooper, J., Coden, A., & Brown, E. (2002). Detecting similar documents using salient terms. Proceedings of the 11th International Conference on Information and Knowledge Management, McLean, VA (pp. 245–251). ACM Press.

Cormack, G., Palmer, C., & Clarke, C. (1998). Efficient construction of large test collections. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia (pp. 282–289). ACM Press.

Fetterly, D., Manasse, M., & Najork, M. (2003). On the evolution of clusters of near-duplicate Web pages. Proceedings of the First Latin American Web Congress, Santiago, Chile (pp. 37–45). Los Alamitos, CA: IEEE Computer Society Press.

Harter, S. (1996). Variations in relevance assessments and the measurement of retrieval effectiveness. Journal of the American Society for Information Science, 47(1), 37–49.

Hawking, D., Voorhees, E., Craswell, N., & Bailey, P. (2000). Overview of TREC-8 Web Track. The 8th Text Retrieval Conference (pp. 131–148). NIST. Retrieved from http://trec.nist.gov/pubs/trec8/papers/web_overview.pdf

Heintze, N. (1996). Scalable document fingerprinting. Proceedings of the 2nd USENIX Electronic Commerce Workshop, Oakland, CA (pp. 191–200).

Hersh, W., Buckley, C., Leone, T.L., & Hickman, D. (1994). OHSUMED: An interactive retrieval evaluation and new large test collection for research. Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland (pp. 192–201). New York: Springer-Verlag.

Hoad, T., & Zobel, J. (2002). Methods for identifying versioned and plagiarised documents. Journal of the American Society for Information Science and Technology, 54(3), 203–215.

Jackson, P., & Moulinier, I. (2002). Natural language processing for online applications: Text retrieval, extraction and categorization. Philadelphia: Benjamins.

Jones, K.S., & van Rijsbergen, C.J. (1975). Report on the need for and provision of an "ideal" information retrieval test collection (British Library Research and Development Report No. 5266). Computer Laboratory, University of Cambridge.

Jones, K.S., & van Rijsbergen, C.J. (1976). Information retrieval test collections. Journal of Documentation, 32(1), 59–75.

Kołcz, A., Chowdhury, A., & Alspector, A. (2004). Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA (pp. 605–610).

Manber, U. (1994). Finding similar files in a large file system. Proceedings of the USENIX Winter 1994 Technical Conference, San Francisco (pp. 1–10).

Marcu, D. (1999). The automatic construction of large-scale corpora for summarization research. Proceedings of the 22nd Annual International

[17] The test collection developed for this project is available for noncommercial research purposes upon request from the first author.

ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA (pp. 137–144). ACM Press.

Miller, C. (1990). Detecting duplicates: A researcher's dream come true. Online, 14(4), 27–34.

Mitchell, T. (1997). Machine learning. New York: McGraw-Hill.

Park, S.-T., Pennock, D., Giles, C.L., & Krovetz, R. (2002). Analysis of lexical signatures for finding lost or related documents. Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland (pp. 11–18). ACM Press.

Rose, T., Stevenson, M., & Whitehead, M. (2002). The Reuters Corpus Volume 1—From yesterday's news to tomorrow's language resources. Proceedings of the 3rd International Conference on Language Resources and Evaluation, Las Palmas de Gran Canaria. Paris: ELDA.

Sanderson, M. (1997). Duplicate detection in the Reuters Collection (Tech. Report TR-1997-5).

Saracevic, T. (1997). Users lost: Reflections on the past, present, future, and limits of information science. ACM SIGIR Forum, 31(2), 16–27.

Schleimer, S., Wilkerson, D.S., & Aiken, A. (2003). Winnowing: Local algorithms for document fingerprinting. Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, San Diego, CA (pp. 76–85). ACM Press.

Shaw, W.M., Jr., Wood, J.B., Wood, R.E., & Tibbo, H.R. (1991). The cystic fibrosis data base: Content and research opportunities. International Journal of Library and Information Science Research, 13, 347–366.

Sheridan, P., Ballerini, J.P., & Schäuble, P. (1996). Building a large multilingual test collection from comparable news documents. Workshop on Cross-Lingual Information Retrieval, Philadelphia (pp. 56–65). ACM Press.

Shrivakumar, N., & García-Molina, H. (1998). Finding near-replicas of documents on the Web. Proceedings of Workshop on Web Databases, Valencia, Spain (pp. 204–212).

Siegel, S., & Castellan, N.J., Jr. (1988). Nonparametric statistics for the behavioral sciences (pp. 284–289). Boston: McGraw-Hill.

Soboroff, I., & Robertson, S. (2003). Building a filtering test collection for TREC 2002. In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto (pp. 243–250). ACM Press.

Tenopir, C., & Cahn, P. (1994). Target & FREESTYLE: DIALOG & Mead join the relevance ranks. Online, 18(3), 31–47.

Turtle, H. (1994). Natural language vs. Boolean query evaluation: A comparison of retrieval performance. Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland (pp. 212–221). New York: Springer-Verlag.

Voorhees, E. (2000). Variations in relevance judgments and the measurement of retrieval effectiveness. Information Processing and Management, 36(5), 697–716.

Voorhees, E., & Harman, D. (2000). Overview of the Sixth Text Retrieval Conference (TREC-6). Information Processing and Management, 36(1), 3–35.

Zhang, Y., Callan, J., & Minka, T. (2002). Novelty and redundancy detection in adaptive filtering. Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland (pp. 81–88). ACM Press.

Appendix A

Sample Contributions From Library Advisory Board Members

1. I believe a definition that relies on a figure like 80% overlap is most desirable since the percentage gives much more precision to the meaning. It would eliminate that case of one article written with 12 different titles in 12 regional publications, all with varying word counts. With stories like those from Dow Jones which are regularly updated, I would, of course, prefer the longer version also presumably with the later time stamp.

2. A definition relying upon 80% accounts for the greatest part of the document, and, perhaps, [include] a disclaimer that you cannot account for the other 20% (which may hold that ever-so-valuable tidbit, for the researcher, that was ignored for inclusion by other sources) [this implies that grouping such similar articles could be useful].

Appendix B

Sample News Articles Illustrating Fuzzy Duplicate

Example A

Dreams Recede As Money Dries Up

This week, **[Luke Casserly]** was sitting in an empty stand at *Marconi's home in* Bossley Park, *Sydney*, pondering how much his star had *fallen* and whether there was still some way to fall. Even his old club, Marconi Stallions, has been slow to rescue him. For a *player* without a contract, the new world is intimidating.

For Casserly, *it has been* sobering. For other Australian players, it is a cautionary tale.

Paris-based Bernie Mandic, the player manager with more experience in these matters than anyone else in the Australian game, paints a bleak picture for Australian players in an environment where the collapse of the pay television industry *overseas* has put professionals *all over the world* out of work.

Example B

Demise of The Goose That Laid The Golden Egg

This week, **[Luke Casserly]** was sitting in an empty stand at Bossley Park, pondering how much his star had *waned* and *wondering* whether there was still some way to fall. Even his old club, Marconi Stallions, has been slow to rescue him *from exile.* For a *footballer* without a contract, the new world order is *an* intimidating *place*.

For Casserly, a sobering *experience.* For other Australian players *on the market*, it is a cautionary tale. *And if they don't believe Casserly, perhaps they should heed the advice of* Bernie Mandic, the player manager with more experience in these matters than anyone else in the Australian game.

Now based in Paris, Mandic paints a bleak picture for Australian players in an environment where the collapse of the pay television industry has put professionals out of work.

Italics indicates terminology differences between the two articles.

DOI: 10.1002/asi