Constructing a Text Corpus for Inexact Duplicate Detection

Jack G. Conrad
Research & Development
Thomson Legal & Regulatory
St. Paul, MN 55123 USA
Jack.G.Conrad@Thomson.com

Cindy P. Schriber
Business & Information News
Thomson–West
St. Paul, MN 55123 USA
Cindy.Schriber@Thomson.com

ABSTRACT

As online document collections continue to expand, both on the Web and in proprietary environments, the need for duplicate detection becomes more critical. The goal of this work is to facilitate (a) investigations into the phenomenon of near duplicates and (b) algorithmic approaches to minimizing its negative 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 inexact duplicate documents.

Categories and Subject Descriptors

H.2.4 [**Information Systems**]: Database Management—*Systems–Textual Databases*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Selection Process*; H.3.m [**Information Storage and Retrieval**]: Miscellaneous—*Test Collections*

General Terms

Experimentation, Measurement, Design, Algorithms

Keywords

test collections, duplicate document detection

1. INTRODUCTION

At Thomson Legal & Regulatory (TLR), massive data environments like Westlaw and Dialog possess on the order of 25 terabytes of data. In such environments, the identification of duplicate documents is an important factor for a practical and robust data delivery platform.

The goal of this work is to apply domain expertise at both the front-end (user representatives) and back-end (professional assessors) of the problem space in order to characterize the duplication existing in large textual collections. We subsequently try to validate the completeness and reliability of this effort with analyses of assessor agreement, error rates, and significance.

The fundamental contribution of this work is the creation of a "deduping" test collection by harnessing:

(a) real user queries;
(b) a massive collection from an operational setting;
(c) professional assessors possessing substantial knowledge of the domain and its clients.

Recent research has often been syntax rather than lexical-based, Web-based (focusing on issues such as URL replication and instability), and offline-based (e.g., examining large numbers of permutations before constructing a feature set). Previous work is thus substantially different than our current efforts that target a dynamic production environment.

2. PREVIOUS WORK

Efforts have been made to construct utilitarian, domain-specific collections that facilitate specific tasks such as multilingual IR, summarization, and filtering. This work appears to be the first to focus on a means of testing "fuzzy" (i.e., inexact or non-identical) duplicate documents while attempting to satisfy expressed user preferences.

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 pages, e.g., [1, 5, 3]. But there has yet to be established a standard IR test collection for duplicate document detection. This was our first necessary step: without a validated test collection, we could not have confidence in the approaches and performance measures that follow.

3. METHODOLOGY

3.1 Background

Initially the Business & Information News (BIN) portion of our organization asked us for technologies to identify and treat duplicate documents. In response, we began characterizing the distribution of duplicate types across our news collections [4] and then proceeded to address the two largest categories of duplicates. At the time, the BIN repository consisted of roughly 55 million news documents.

3.2 Problem Definition and User Feedback

We began by conducting a feedback session with 25 members of our Library Advisory Board who represented high-level users from our clients' enterprises and firms. Most of the group's formal training is in the field of Library Science. As such, these individuals are uniquely positioned to provide domain expertise in their focus areas and are an excellent group to consult. In all, 17 of the 25 participants provided non-trivial replies to our suite of questions.

The objective of the feedback session was to describe, both qualitatively and quantitatively, the nature of the most annoying duplicate documents. This exercise resulted in the following description: a non-identical 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% terminology overlap "duplicates" would be problematic. These guidelines produced a working definition of "near duplicate" documents with which we proceeded.

3.3 Corpus Generation and 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. 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

SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK.
ACM 1-58113-881-4/04/0007.

natural language search capability, depending on the preference of the user [6]. After running these queries against the ALLNEWSPLUS database consisting of approximately 45 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, in order to identify their duplicate sets. This process produced standard training and test sets against which computational approaches would be compared.[1]

3.3.1 Details of Document Inspections

In this trial, we applied the definition of inexact duplicate that was generated by a customer user group described in Section 3.2. To formally review the duplication status of the result sets, we assembled two teams of two assessors consisting of client research advisors. The 100 queries were divided into two sets of 50, the first set to be used to train the system and the second set to test it. The process by which the query results were judged was scheduled over four weeks time. 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.

Week 2 served as an arbitration week. When members of the same team disagreed about a duplicate set, a member of the other 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 three.

Table 1 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 4, 5, or 6 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).

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

Table 1: Distribution of Total Resulting Duplicate Sets

3.4 Inter-assessor Agreement

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.

We used the Kappa statistic for nominally scaled data in order to compare our inter-assessor concordances before arbitration over the 100 result sets [2]. 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 2.

[1] "Training" is not used here in the Machine Learning sense. It signifies an initial round to set an algorithm's optimal parameters.

Given a result set of $n = 20$ docs, 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 non-duplicate. We computed the Kappa statistic over the comparison space described.

Computational linguists have taken $\kappa = 0.8$ as the norm for significantly good agreement, although some argue that there is insufficient evidence to choose 0.8 over, for instance, other values between 0.6 and 0.9.

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 Qrys) *(Teams A & B)*	$\kappa = 0.7829$ (0.7784)	

Table 2: Kappa Statistics for Inter-assessor Agreements for Duplicate Set Identification [macro-averaged scores (micro-averaged scores in parens)]

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, $z = \kappa/\sqrt{var(\kappa)}$ in order to test the hypothesis $H_o : \kappa = 0$ against the hypothesis $H_1 : \kappa > 0$ [2].

The above 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). Worth underscoring is that this agreement level occurs before the arbitration rounds.

4. SUMMARY & CONCLUSIONS

The growth of electronic data environments has expanded the need for various forms of duplicate document detection. Our exploration addresses a real world replication problem in the news domain. Our methodology invited library scientists or meta-level users to define the scope of the problem, and commissioned two teams of searchers to use the working definition and principled methods to identify non-identical duplicates in the resultant corpus. We have also tried to validate the decisions of our assessors using a Kappa analysis. For the inexact duplicate detection task, our applied test collection proved beneficial; follow-up trials have uncovered feature sets that can serve as strong indicators of degree of duplication. Details of subsequent deployments of the corpus can be found in [4].

5. REFERENCES

[1] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig. Syntactic clustering of the web. In *Proceedings of WWW6 '97*, pages 391–404. Elsevier Science, April 1997.

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

[3] A. Chowdhury, O. Frieder, D. Grossman, and M. C. McCabe. Collection statistics for fast duplicate document detection. *ACM TOIS*, 20(2):171–191, April 2002.

[4] J. G. Conrad, X. S. Guo, and C. P. Schriber. Online duplicate document detection: Signature reliability in a dynamic retrieval environment. In *Proceedings of CIKM'03*, pages 443–452. ACM Press, Nov. 2003.

[5] N. Shrivakumar and H. García-Molina. Finding near-replicas of documents on the Web. In *Proceedings of Workshop on WebDB '98*, pages 204–212, March 1998.

[6] H. Turtle. Natural language vs. Boolean query evaluation: A comparison of retrieval performance. In *Proceedings of SIGIR '94*, pages 212–221. Springer-Verlag, July 1994.