An Exploratory Study of the W3C Mailing List Test Collection for Retrieval of Emails with Pro/Con Arguments

Yejun Wu & Douglas W. Oard
College of Information Studies and
Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742
{wuyj,oard}@glue.umd.edu

Ian Soboroff
National Institute of Standards and Technology
Gaithersburg, MD 20899
ian.soboroff@nist.gov

ABSTRACT

The W3C mailing list test collection, an information retrieval test collection for email, was developed for the TREC Enterprise Search track in 2005. One task in that track was to retrieve emails that contribute at least one pro/con related to a specific topic. This paper describes the test collection and presents a preliminary evaluation of its suitability for evaluating such systems, including an analysis of topic types found in the collection, characterization of inter-assessor agreement on pro/con judgments, and an example of the evaluation results that can be obtained using the collection. There is clear evidence that the collection is useful in its present form, but several areas for improvement can be identified. In particular, some topic types found in the collection do not seem well suited to pro/con judgment. The paper concludes with suggestions for future work on the design of test collections and information retrieval systems for this task.

1. INTRODUCTION

Informal genres such as email pose new challenges for the design of information retrieval systems that help people to find information that they seek. Information retrieval research has traditionally focused on relatively formal genres (e.g., news stories) in which topic and (more recently) source authority are typically seen as important. Informal genres such as personal letters, discussion boards, mailing lists, emails, Usenet, weblogs, and spoken word collections offer more scope for incorporating additional search criteria, including expressions of sentiments such as opinions or attitudes. It is easy to envision numerous situations in which an ability to characterize sentiment would be useful: political campaigns may wish to know public opinions and attitudes about a candidate, companies may want to know their customers' views about their products, and decision makers may wish to understand the opinions and attitudes of key stakeholders. A study of 17 genres selected from the British National Corpus and other sources indicated that emails are significantly less formal than news and weblogs, and that they are similar in formality to personal letters [18]. This suggests that email would be a good starting point from which to examine the practical utility of sentiment detection in information retrieval systems that are designed to work with informal media.

CEAS 2006 Third Conference on Email and AntiSpam, July 27-28, 2006, Mountain View, California, USA.

Opinion classification in documents has been studied in the domains of movie and product reviews [21]. Reviews collected on well known Web sites (e.g., Rottentomatoes.com, Amazon.com, and C|net) often include both text and ratings, making such studies easy to construct [4], although the utility of the results are open to question (since reviews without ratings may be rare, and systems trained on reviews may not generalize well to other tasks). Opinion detection in news articles [1, 30] and customer emails [5] has also been studied, but we are not aware of prior work on opinion or attitude detection in mailing lists, Usenet news, or similar sources.

Mailing lists are sometimes public, and public mailing lists are sometimes archived. Moreover, mailing lists can sometimes serve as important sources of institutional memory, making it fairly straightforward to identify (or at least envision) realistic information needs. They therefore represent a practical basis for constructing an information retrieval test collection. In 2005, the Text Retrieval Conference's (TREC) Enterprise Search track (TRECENT) therefore began the process of developing what we believe to be the world's first test collection for evaluating topic-oriented search of a relatively large email collection. In this paper, we focus on one task for which that test collection was designed: identifying emails that contribute at least one statement in favor of or against a specified topic in new (not quoted) text (that is, identifying at least one pro or con argument about the topic).[1]

An information retrieval test collection typically includes a set of documents, a set of information needs (called *topics* in TREC), and *relevance judgments*, an indication of which documents should be retrieved in response to which topics [29]. A *document* in the TRECENT 2005 collection was defined to be an individual email message. The *topic* statements consisted of three sections: a unique identifier (used only for accounting purposes), a "title" field (representing what a searcher might type in as a terse initial query), and "narrative" field containing a concise statement of the criteria for assessing topical relevance (see the next section for examples) [29]. Two types of relevance were defined: topical (also called "partial") relevance, and pro/con (also called "full") relevance. For a document to be pro/con (or fully) relevant, it must be topically relevant and it must additionally present at least one pro or con argument in new text (i.e., not in text that was simply repeated from an earlier

[1]The TRECENT guidelines, available at http://www.ins.cwi.nl/projects/trec-ent/wiki/, also describe the other tasks.

email in the same thread) about the topic.

While the absolute value of a retrieval effectiveness measure can give some sense for how well a system supports the retrieval task, judging relevance is necessarily a somewhat subjective process, and differences in judgments will yield different values for any effectiveness measure. The typical way in which these collections are used, therefore, is to hold the judgments constant, and to rank systems in decreasing order of retrieval effectiveness with that same set of judgments. In other words, it is not the absolute value of the measures where the information retrieval community focuses, but rather on the preference order among the systems. A test collection is useful if it yields a preference order that is stable across users—in other words, we desire that a set of retrieval systems be scored consistently relative to each other, even if the scores assigned to each system vary systematically with the strictness of the judgments provided by one assessor or another.

Each search engine returns a ranked list of documents in response to each query. One widely reported measure is *Average Precision* (AP). AP is defined (for a single topic) as the expected value of the precision (the density of relevant documents), with the expectation computed over the set of documents that were judged to be relevant. This is intended to model the satisfaction of a searcher who begins to scan a list of documents that are ranked in some approximation of a best-first order from the top and stops after having seen some desired (but unknown) number of relevant documents. AP values vary markedly from one topic to another, and it is not uncommon for the preference order between systems to be different for different topics. Since we don't know what topic the searcher will ask about next, it is common to report the expectation over a suite of (at least 40) representative topics of AP as a measure of expected retrieval effectiveness on future topics; this measure is known as *Mean Average Precision* (MAP) [29].

Among the participants in TRECENT 2005, Zhu et al. [33] were the only ones to specifically look at the pro/con task. They trained a Support Vector Machine (SVM) using a small set of hand-tagged training data, but other teams achieved far higher MAP with the pro/con judgments by using only topical (i.e., word-overlap) search techniques. Now that the full test collection is available, we are able to take a more comprehensive look at the suitability of the test collection for that task. We begin by a brief review of previous studies on opinion classification, followed by a description of the test collection, an analysis of the topic type distribution, and an analysis of inter-assessor agreement on individual judgments and the stability of system preference order across judgment sets from different assessors. We then describe the implementation of a retrieval system trained using held out data from the test collection (in a round robin fashion), and report its performance by topic type. We conclude with suggestions for future work on topic selection, assessor training, and information retrieval system design.

2. RELATED WORK

The essential issues in document-level sentiment analysis (or opinion classification) are to identify how sentiments are expressed in texts and whether the expressions indicate positive (favorable, recommended) or negative (unfavorable, not recommended) opinions toward the subject [17]. Both linguistic and machine learning approaches have been applied.

There are many studies in classifying movie and product reviews. Turney [27] applies a linguistic approach to classify Epinions reviews. Pang et al. [21] apply classical text classification techniques (i.e., Naive Bayes, maximum entropy, and SVM) to the task of classifying movie reviews as positive or negative. Kennedy and Inkpen [11] use three types of valence shifters (i.e., negations, intensifiers and diminishers) and machine learning algorithms to detect the semantic orientation of movie reviews. More sentiment classification for reviews of products, movies, paper peer reviewers, and stock investment are discussed in [3], [9], [15], and [16].

Durbin et al. [5] construct a system for affect rating of texts with a particular domain (i.e., customer emails) with multiple languages with modest effort. The system takes 5 steps. First, a few thousand sentiment words are manually collected and rated by a number of raters. Second, a document is tagged with a part-of-speech tagger, and individual rated words (i.e., those on a proprietary list) are identified. Third, valence modifiers such as "very" or "slightly" are detected, and the ratings of words immediately following are recalculated via a modification function. Fourth, for sentences containing both rated words and negation, syntactic rules are applied to determine whether the negation applies to the rated words or not, and the sentence ratings are adjusted. Finally, an overall rating is assigned to the document. Evaluated on the collection of movie reviews assembled by Pang et al. [21], the system has an accuracy of 63%, well below the best result of 83% obtained by Pang et al. [21] who train various machine learning algorithms on that specific data set.

Beyond classifying a document into positive and negative categories, Pang and Lee [22] classify movie reviews with respect to a multi-point scale (e.g., one to five "stars"). Wilson and Wiebe have also classified the strength of opinions [30] and contextual polarity of the polar expressions [31].

The previous studies on opinion classification have been conducted in the domains of movie and product reviews, stock investment comments, customer emails, and news articles with relatively small data sets (typically hundreds to several thousand documents). In the next section we introduce the W3C mailing lists test collection for retrieval of emails with pro/con arguments.

3. THE W3C TEST COLLECTION

3.1 Description of the Test Collection

The W3C mailing list test collection is a large public test collection for evaluation of content-based email search using World Wide Web Consortium (W3C) mailing lists. The W3C mailing list collection was crawled from w3c.org in June 2004; the National Institute of Standards and Technology (NIST) distributed the collection to the TRECENT participants. Each message in the collection was embedded in a Web page with extensive XML markup (usually generated by the hypermail utility program) to format the most important fields from the message for display to end users. We used a Java SAX parser to recover the original RFC-822 header structure and to extract the body text. For messages that the SAX parser failed to parse (such as X-Mail messages), additional processing using Perl scripts was performed. This yielded 174,311 messages with a total size of 515 MB.

TRECENT 2005 had three tasks - Expert Search, Known

Item Search and Discussion Search (DS). We focus on the DS task here. DS is a conventional ad hoc retrieval task in which the user is searching for arguments in favor of or against some point in an email archive. This might be used, for example, to assemble design rationale when considering a change to a previously published W3C standards document. A *pro/con (or fully) relevant* document in the DS task is an email message that contributes at least one pro or con related to the specified topic in new (not quoted) text; a *topically (or partially) relevant* document is an email message that addresses the specified topic in new (not quoted) text but provides no pro/con related to the specified topic in new text.

The participating teams performed topic development and relevance judgment. A total of 60 DS topics were created, one of which was subsequently removed from the collection because no relevant documents were found. No training topics for the DS task were available since the test collection had not been created yet. Each DS topic was assigned to two participating groups to judge relevance in order to support computation of inter-assessor agreement. The top 50 retrieved documents from the four highest priority runs from each group were pooled for relevance judgement. There were an average of 529 messages per topic across the 59 pools (ranging from 249 to 865). Next we examine the types of these 59 topics [2].

3.2 Emergent Categories of Topics

Since an information retrieval test collection includes documents, topics, and relevance judgments, and we want to do some preliminary evaluation of the test collection for the pro/con retrieval task, we begin by a topic type analysis followed by an analysis of inter-assessor agreement of the relevance judgments.

The goal of the topic type categorization is to investigate what types of topics have been developed, and to identify topic categories that may be more amenable to pro/con classification. We did this by manually examining the 59 TRECENT DS topics to identify emergent categories. An example topic is listed under each category. Note that topics may be classified into multiple categories.

- A: Comparison, usefulness, relationships, etc.
  A0: Comparison among design options or standards (pro/con, advantage/disadvantage, whether/not) (2 topics in mutual categories).
  A1: Design compliance (or conflict) with standards.
  A2: usefulness, feasibility or suitability of a design
  A3: relationship among issues.

  Example: Topic 55 (in category A0 and B below)
  Title: Browser technology support incompatibility
  Narrative: A relevant message will offer possible solutions to the problem of different browsers supporting different technologies (e.g., scripting languages), and show advantages/disadvantages of these solutions.

- B: Method, tip, solution (How to use X, how to solve problem, how to fix bug)

  Example: Topic 19
  Title: abbreviations and acronym expansion
  Narrative: A relevant message discusses methods used to expand acronyms and issues people need to be aware of in this area.

- C: Discuss an issue/policy

  Example: Topic 27
  Title: P3P English translation
  Narrative: Relevant messages will discuss the translation of the P3P element definitions into plain English.

- D: Problems, Impact, etc.
  D0: Problems/bugs/vulnerabilities of designs.
  D1: Impact/Consequence/Effect of design policies (1 topic in mutual category).

  Example: Topic 44
  Title: Shared key authentication
  Narrative: A message should discuss the effects of the usage of shared secret authentication in TLS.

- E: Definition, functionality(what is X, what is the use of X)

  Example: Topic 40
  Title: rational definition of identifier
  Narrative: A relevant message will discuss what is an appropriate definition of identifier.

- F: Reason, design rationale (Why is X)

  Example: Topic 1
  Title: if-else in xslt
  Narrative: A relevant message will discuss the reasons for the non-availability of an if-else construct in xslt.

Category	Number of Topics
A	29
B	11
C	11
D	4
E	3
F	1

Table 1: Number of topics in each category.

The topics are not equally distributed across the categories. As shown in Table 1, about half of the topics are related to A, about 1/6 to B, about 1/6 to C, and nearly 1/6 to the others. We analyze the inter-assessor agreement of relevance judgments by category in the following subsection.

3.3 Inter-assessor Agreement of Judgments

Inter-assessor agreement of relevance judgments is the agreement between primary and secondary assessors on whether a document is fully relevant, partially relevant, or not relevant at all. TRECENT had two groups of assessors: primary assessors who developed the topics and secondary assessors who were other participants. Here we examine the inter-assessor agreement on all the pooled documents of the 48 topics for which relevance judgments exist from both groups

of assessors. Measuring inter-assessor agreement allows us not only to examine the quality of relevance judgments *per se* and hence the stability of system preference order across judgment sets from different assessors, but also to identify topic types or topics that are possibly ill-defined for pro/con retrieval.

Here we introduce three measures that are related to inter-assessor agreement of relevance judgements—agreement overlap, Cohen's Kappa, and Kendall's tau. Overlap and Cohen's Kappa directly measure inter-assessor agreement, while Kendall's tau quantifies the effect of disagreements on relative ranking of systems using the test collection.

Both agreement overlap and Cohen's kappa have been used to quantify the amount of agreement among different sets of relevance judgments [12, 19]. Overlap is defined as the size of the intersection of the relevant document sets divided by the size of the union of the relevant document sets [28]. Since two assessors can reach an agreement by chance, we introduce Cohen's kappa which is a chance-corrected measure of agreement [8]. For the inter-assessor agreement on topical relevance judgments, topic-averaged overlap is 0.29 (with a possible range of 0 to 1) and overall kappa is 0.42 (with a possible range of -1 to 1). The correlation between overlap and kappa is 0.9661, which is statistically significant at $p < 0.01$. For the agreements on pro/con relevance (i.e., full relevance vs. not full relevance) judgments, topic-averaged overlap is 0.275 and overall kappa is 0.3. The correlation between overlap and kappa is 0.9787, which is statistically significant at $p < 0.01$. The statistical significance of the correlation scores in both cases indicates that as an inter-assessor agreement measure overlap is as good as kappa for this test collection.

The inter-assessor overlap for flat text retrieval is typically between 0.4 and 0.5 [23, 28], and the topic-averaged overlap for a spontaneous speech test collection is 0.44 [19]. For the W3C test collection, both overlap scores are below 0.3, which is relatively low.

Often in information retrieval there is a low level of inter-assessor agreement of relevance judgments (because relevance is subjective and idiosyncratic) but the disagreements about relevance don't affect the relative measurement of system A being better or worse than system B. In other words, a low level of agreement between two assessors doesn't necessarily mean that systems would be ranked differently if they were scored using the relevance judgments of each assessor. One of the reasons system rankings remain stable despite inevitable marked differences in the relevance judgments is that evaluation results (such as MAP) are reported as averages over many queries [12]. While evaluation results for individual topics (or queries) can vary widely, an average over a sufficient number (such as more than 40) of queries is more stable [28].

The *effect* of inter-assessor disagreement on evaluation results is commonly measured using Kendall's tau correlation. In other words, Kendall's tau measures the *effect* of disagreement, rather than the *amount* of disagreement itself. By comparing the rankings of systems according to the primary and secondary judgments, Kendall's tau calculates the distance between two rankings of retrieved items as the minimum number of pairwise adjacent swaps to turn one ranking into the other, and the distance is normalized by the number of items being ranked such that the correlation is between -1.0 (for the correlation between a ranking with its perfect reverse) and 1.0 (for the correlation of two identical rankings) [28].

Type	Topics	Ovlp-p	Ovlp-t	Kap-p	Kap-t	Kap-3
A	26	0.27	0.32	0.37	0.36	0.29
A0	18	0.26	0.35	0.38	0.37	0.29
A1	2	0.03	0.08	0	0.02	0
A2	4	0.40	0.36	0.46	0.46	0.4
A3	2	0.29	0.29	0.38	0.43	0.35
B	10	0.23	0.32	0.26	0.33	0.22
C	8	0.41	0.47	0.38	0.38	0.31
D	4	0.25	0.31	0.49	0.50	0.41
D0	1	0	0.11	0	0.19	0.09
D1	3	0.33	0.38	0.50	0.50	0.41
E	2	0.04	0.34	0	0.54	0.24
F	1	0.44	0.32	0.61	0.47	0.35
All	48	0.275	0.29	0.30	0.42	0.29

Table 2: Inter-assessor agreement on 48 topics by topic types

*Note: three topics are classified into 2 categories, so the number of topics in the table does not directly sum up to 48. Ovlp-p is overlap of agreement on pro/con (full) relevance, Ovlp-t is overlap of agreement on topical (partial) relevance, Kap-p is kappa for pro/con relevance, and Kap-t is kappa for topical relevance. Kap-3 is kappa across 3 categories - fully relevant, partially relevant, and irrelevant.

P Values	Ovlp-p	Ovlp-t	Kap-p	Kap-t	Kap-3
p@5	0.21	0.69	0.21	0.70	0.32
p@9	0.20	0.73	0.17	0.68	0.39

Table 3: p values of ANOVA with the inter-assessor agreement as a dependent variable and the topic type as an independent variable

*Note: p@5 is significance value when topic type has 5 levels: A, B, C, D, and E. p@9 is significance value when topic type has 9 levels: A0, A1, A2, A3, B, C, D0, D1, E. Other notations are the same as Table 2.

For the DS task, Craswell et al. [2] computed Kendall's tau for the effect of inter-assessor disagreement for topical relevance judgments on evaluation results using the 48 topics. The tau correlation between rankings based on the two sets of judgments was 0.763 [2], which is significant although not as strong as in other test collections [28]. We also computed Kendall's tau for the effect of inter-assessor disagreement on full relevance from both judgments, which is 0.776, very close to the tau for topical relevance judgements. Common practice in the retrieval community is to consider tau values larger than 0.9 as essentially identical, and less than that as possibly indicating important differences, but this rule of thumb is not formally grounded.

Note that tau is lower than we'd like it to be, so we should be paying attention to the overlap and kappa scores to identify possibly ill-defined topics or topic types in order to get that tau up to the level of a reliable collection.

There are a few factors which probably affect inter-assessor agreement - topic, assessor pair, and topic type. The first

Factors	Ovlp-p	Ovlp-t	Kap-p	Kap-t	Kap-3
Type@5	0.005	0.158	0.008	0.220	0.045
Topic	0.704	0.430	0.404	0.464	0.503
Interaction	0.000	0.002	0.000	0.005	0.001
Topic@9	0.010	0.378	0.017	0.434	0.128
Topic	0.467	0.417	0.343	0.420	0.413
Interaction	0.000	0.020	0.001	0.039	0.009

Table 4: p values of the factors of two way ANOVA with the inter-assessor agreement as a dependent variable, and the topic type and topic as factors
*Note: Type@5 is the topic type factor with 5 levels: A, B, C, D, and E. Type@9 is the topic type factor with 9 levels: A0, A1, A2, A3, B, C, D0, D1, E. Interaction is the interaction between the two factors. Other notations are the same as Table 2.

is topic itself: some topics are harder to judge relevance of their documents than others. As seen in the examples shown in section 3.3, topic 19 and 27 are vaguely defined for their pro/con requirements and an inter-assessor agreement analysis shows that both of their full relevance agreement overlap scores are below 0.2 (see Table 7 in Appendix).

Another factor is the assessor pair: some assessor pairs may have higher agreement than others. In TRECENT 2005, this is hard to work with since we do not know which individual person at each research site judged which topic. We wondered whether there were systematic differences in labeling full/partial relevance between primary and secondary assessors by comparing the two assessor group's fractions of relevant emails which are fully relevant. However, a t-test showed that there is no statistically significant difference between the two fractions (p = 0.42, two-tailed test).

Here we focus on the factor of topic type. An inter-assessor agreement analysis by category shows that A2, C, D1, and F type topics have both relatively high overlap and kappa scores, as shown in Table 2. An analysis of variance (ANOVA) with the inter-assessor agreement as a dependent variable and the topic type as independent variable shows that the agreement difference between topic types is not statistically significant (see Table 3). However, Table 3 also shows that the inter-assessor agreements on full relevance across topic types are much more likely to be different than the agreements on topical relevance across topic types. We wondered whether the fractions of relevant emails which were fully relevant were different between topics; if yes, it might indicate quantitatively that certain types of topics do not lend to pro/con discussion. However, a series of t-tests on the difference of the means of the fractions between topic types show no statistical significance (significance values range from 0.34 for A3 to 0.95 for E, two-tailed tests).

In order to examine whether the *topic type* factor is more significant than the *topic* factor, a series of two way ANOVA tests are performed with the inter-assessor agreement as a dependent variable, and topic and topic type as factors (see Table 4). Here topic type has 5 or 9 type levels (see Table 4) whereas topic has 3 *topic difficulty* levels: *easy* topics, *hard* topics, and *unsure* topics. *Easy* topics are the 27 topics on which our pro/con retrieval system (described in the next section) achieves improvements of performance over our baseline system. *Hard* topics are the 16 topics on which our pro/con retrieval system performs worse than the baseline. *Unsure* topics are the 7 topics that are not evaluated with our pro/con retrieval system for reasons described in the next section. As shown in Table 4, topic type has a statistically significant effect on the inter-assessor agreement for full relevance (but not topical relevance) judgments at $p < 0.05$ whereas topic has not, and neither topic nor topic type has a significant difference in agreement for topical relevance judgments. The significant effect of interaction between topic type and topic on the agreement of full relevance judgments at $p < 0.05$ tells us that the effect of one variable depends on the level of the other variable.

In this section, we have developed topic types (or categories) and analyzed the inter-assessor agreement within each category. We find that some categories have more of a pro/con nature, and our pro/con retrieval system which we describe next finds them. We also find that inter-assessor agreement is somehow different across topic categories (although not statistically significantly different), and that topic category is a significant factor of inter-assessor agreement of full relevance judgments whereas topic is not, so we offer suggestions in section 5 about how to better design a pro/con email test collection.

4. TOPIC AND PRO/CON INFORMATION RETRIEVAL

4.1 Pilot Study

Inspired by OASYS [1] which applies a manually collected lexicon of opinion and attitude words to analyze opinions in news articles, we conduct a pilot study for topic and pro/con retrieval in the same spirit. We manually process the first 200 documents (by document id sorting) in the official query-relevance set (that is, the *qrels* file, the concatenation of one relevance judgment set per topic), which includes both non-relevant and relevant (i.e., both partially and fully relevant) documents, then extract about 50 words or phrases that are most obvious (to the first author) to be usually used to express pro/con arguments. INQUERY (Version 3.1p1 for Solaris) is used as the search engine. The pro/con words and phrases are fed into INQUERY's weighted sum (wsum) operator in query formulation. Each topic term is assigned a weight of 1.0 and each pro/con term gets a weight varying from 0.01 to 0.5. With all opinion/attitude term weights assigned to 0.05, our best pilot system achieves a MAP score of 0.3042 for 59 topics, a minor improvement of 2% over the best system TitleTrans (with a MAP of 0.2969) for the TRECENT 2005 DS task [2]. The improvement is due to the fact that the system retrieves fully relevant emails earlier than before, thus increasing the precision.

Here are some example pro/con words (or phrases) used in the pilot system: pro, con, agree, disagree, advantage, disadvantage, strength, weakness, shortcoming, limitation, downside, vote for, veto, dislike, incorrect, correct, wrong, pointless, useless, positive, negative, argue, doubt, suspect, guess, understand, insane, handy, reasonable, convenient, inconvenient, annoying, unhappy, make sense, two cents, mistake, etc. They are either nouns, verbs, or adjectives.

The pilot system is actually developed with the notion of query expansion using relevance feedback. Our positive result in the pilot study encourages us to employ a systematic approach to learn words that are used to express pros/cons in the W3C mailing list test collection. This leads to our

system design described next.

4.2 System Architecture

Not all the 59 topics are appropriate for training and evaluation. Topics 5, 6, 13, 18, 23, 35, and 57 have been excluded since fewer than 5 pro/con documents are found in their official query-relevance set (i.e., the qrels file). When there are few relevant documents the average precision measure is unstable in that small perturbations in the document ranking can cause large differences in the average precision [28]. Topics 1, 9, and 46 have also been excluded since no *partially* relevant documents are found in their official qrels. This leaves us a total of 49 topics to work with. Note that these 49 topics are not all the same as the 48 topics discussed in the previous section of inter-assessor agreement because both are subsets of the 60 topics. The two subsets serve different analysis purposes.

We apply separate models for retrieving on-topic emails without pro/con arguments related to the topics and for retrieving on-topic emails with pro/con arguments related to the topics, then combine the two lists of retrieved emails to generate a final ranked list of emails. Since INQUERY's wsum operator combines two ranked lists with a weighted sum of relevance scores, we employ INQUERY as our back end system. During query processing, INQUERY removes stop words and stems the query terms.

First, the email messages are indexed in their original form (i.e., no suppression of quoted text) using INQUERY. Although the goal is to retrieve emails with pro/con arguments in new (not quoted) text, we do not exclude quoted text from the emails since earlier studies showed that removing quoted text was harmful when searching single emails [14].

Second, for each topic, an INQUERY query is formulated using the "title" (or "query") fields only. Then the index is searched with this set of 49 queries to retrieve a ranked list of emails for each topic. This is our baseline system which is not specifically aimed to retrieve on-topic emails with pro/con arguments related to the topics, but is evaluated of its performance at the full relevance level. The MAP score of the baseline for pro/con retrieval is 0.2743 across 49 topics, very close to TitleTrans which is the best system for topic and pro/con retrieval of TRECENT 2005 [2] (with a MAP of 0.2765). Here TitleTrans' MAP across 49 topics is computed using the same techniques introduced in [14] in order to make the two MAP scores comparable.

Third, for each topic, the fully relevant and topically relevant documents of the remaining 48 topics are used as training data to extract a pro/con feature vector which is then used to formulate an INQUERY query to search the index. The way we extract the pro/con features is described in the next subsection. This retrieves a ranked list of emails with pro/con arguments for each topic. Note that the pro/con arguments are not necessarily related to the specified topics.

Finally, for each topic, the two ranked lists are combined by a weighted sum of relevance scores to get a final ranked list.

4.3 Round Robin Experiment Design

For each topic, a pro/con model profile is trained in a round-robin fashion (i.e., leave-one-out or k-fold cross-validation with k = 48) using the remaining 48 topics in order to assess the model performance. The profile is first learned within a topic from both its positive and negative documents aiming at filtering out topical relevance effects, then is combined across all the 48 topics to get a feature vector. MAP is applied to measure the effectiveness of the learned model.

4.4 A Rocchio-style Implementation of the System

Since our goal is to retrieve documents not only on topic but also having pro/con arguments related to the topics, the Rocchio method [24], an information retrieval approach based on query expansion using relevance feedback is a good fit here. Another reason we use this method is that a classifier built with this method is a baseline classifier in the text classification domain; so if a baseline classifier works well for pro/con classification, the test collection is apparently useful, and we will be confident that a better classifier will do a better job. The purpose of implementing such a system is to explore the utility of the test collection.

The Rocchio method is used for inducing linear, profile-style classifiers and it is perhaps the only text classification method rooted in the information retrieval tradition rather than in machine learning [26]. Applied to text classification, it produces a prototype vector for each class as a weighted average of positive and negative training examples [10, 13]. The Rocchio method computes an expanded query as: [7]

$$Q_1 = Q_0 + (\beta \sum_{i=1}^{n_1} \frac{R_i}{n_1} - \gamma \sum_{i=1}^{n_2} \frac{S_i}{n_2})$$

Here Q_0 and Q_1 are the vectors for the initial and expanded queries correspondingly. R_i and S_i are the vectors for the positive (relevant) document i and negative (non-relevant) document i correspondingly. n_1 and n_2 are the number of positive and negative documents chosen correspondingly. β and γ are the parameters for tuning the importance of positive and negative documents. It is fairly common to set β to 1 and γ to 0, in which case the profile (the parenthetical clause) becomes simply the centroid of the positive training examples. For our experiments, we reinterpret the profile as a term vector that is biased in favor of pro/con arguments that is computed as a log-odds ratio.

For each topic, we have 48 topics for training, and 1 topic (i.e., this topic) for evaluation. Basically, for each topic, we want to learn a vector of pro/con features (i.e., words) from the positive (fully relevant) and negative (topically relevant) documents of the 48 training topics. Specifically a feature vector is first learned from both the positive and negative documents within a topic (in an attempt to remove the topical relevance effects), then the vector is accumulated through all the 48 training topics.

The training data (i.e., positive and negative documents) are not equally distributed across the 49 topics. We think that the topics having more training data are better topics, so we assign a topic weight to each topic i, which is defined as:

$$Topic_i Weight = log[min(Npos_i + 1, Nneg_i + 1)]$$

Here each topic i has $Npos_i$ positive documents and $Nneg_i$ negative documents, and has a topic weight computed as the logarithm of the minimum of $Npos + 1$ and $Nneg + 1$.

Our pro/con feature vector is then selected with the logarithm of odds ratio function [6] weighted by the topic weight:

Category	Topics	Baseline MAP	Pro/con MAP
A	27	0.2736	0.2907
B	10	0.2838	0.2859
C	9	0.2293	0.2272
D	3	0.2417	0.2490
E	3	0.2864	0.3013

Table 5: Comparison of MAP by category.

$$\sum_{i=1}^{48}(Topic_i Weight * log\frac{\frac{TFpos_i+1}{Npos_i}}{\frac{TFneg_i+1}{Nneg_i}})$$

Here $TFpos_i$ is the term frequency calculated from the positive documents of topic i with stop words excluded, and $TFneg_i$ is the term frequency calculated from the negative documents of topic i with stop words excluded. For each topic, we compute the logarithm of odds ratio weighted by the topic weight as the term weight. Accumulating the term weights across the 48 training topics, we get a grand list of words sorted by their term weights. In our experiments, we select the top n terms and then construct an INQUERY structured query in the following way:

#wsum(1.0 w_1 #wsum (w_1 $\frac{w_1}{m}$ T_1 $\frac{w_1}{m}$ T_2 ... $\frac{w_1}{m}$ T_m)
w_2 #wsum(w_2 $\frac{w_2}{n}$ PT_1 $\frac{w_2}{n}$ PT_2 ... $\frac{w_2}{n}$ PT_n)

Here T_1, T_2, and T_m are topic terms for topic i; PT_1, PT_2 and PT_n are pro/con terms for topic i; w_1 and w_2 are the weights assigned to the blocks of topic terms and pro/con terms correspondingly ($w_1 + w_2 = 1$), m is the total number of topic terms for topic i, and n is the total number of pro/con terms we select. The whole idea of the query is to assign each topic term a weight of $\frac{w_1}{m}$, and each pro/con term a weight of $\frac{w2}{n}$. We have tested w_1 with 0.3, 0.4, 0.5, 0.6, and 0.8, and n with 50, 80, 100, 200, 300, 500, and 1000. Our best result (reported in Section 4.5) comes with $w_1 = 0.3$, $w_2 = 0.7$, and $n = 100$. It is important to recognize that these paramaters were chosen based on results from the same test collection; this is the best we can do until a separate development test set becomes available.

4.5 Results

Our topic and pro/con retrieval system achieves a MAP score of 0.2857 for 49 topics, a 4.2% improvement over our baseline (with a MAP score of 0.2743), and a 3.3% improvement over TitleTrans (the best system of TRENCENT 2005, with a MAP score of 0.2765). Both our pro/con system and baseline are evaluated with the same query-relevance set (qrels) at the full relevance level. A Wilcoxon signed-rank test for paired samples shows that the difference of MAP scores between our pro/con retrieval system and our baseline is statistically significant at $p < 0.05$, and that the difference of MAP scores between our system and TitleTrans is marginally statistically significant at $p = 0.05$.

A comparison between our topic and pro/con retrieval system and the baseline by category (as shown in Table 5) reveals that major MAP improvements happen in the A and E categories. Table 6 shows the number of topics for which our system's MAP increases or decreases (compared with the baseline) by category. Category B and C have a half

Category	Topics	Topics MAP Up	Topics MAP Down
A	27	18	9
B	10	5	5
C	9	5	4
D	3	2	1
E	3	3	0

Table 6: Number of topics for which Rocchio pro/con system's MAP increased or decreased by category.

of the topics for which our system's performance improves, and the other half decreases. Category A have 2/3 of the topics for which the our system's performance improves, and the other 1/3 decreases. This suggests that some topics in the B and C categories may be inappropriate for pro/con retrieval, or need to be better defined for pro/con relevance judgment, and topics in category A and E are largely better defined for pro/con retrieval. Interestingly all the three topics in category E have shown MAP improvements, and we wondered whether that this might have happened by chance. A t-test on category E indicates some evidence that the two sets of MAP scores are drawn from different distributions, with only an 11% chance that the distributions are the same (two-tailed test). The t-test has proven to be a reliable basis for comparing the effectiveness of two information retrieval systems [25] even though the assumption of normality is not necessarily satisfied.

5. CONCLUSION AND FUTURE WORK

We have done an exploratory evaluation of the W3C mailing list test collection through a topic type analysis and an inter-assessor agreement analysis within each topic category. With that background, we then directly explored the utility of the collection by using it to evaluate a topic and pro/con retrieval system. Here we describe our conclusion and future work for those two issues.

5.1 Test Collection Evaluation and Design

The inter-assessor agreement analysis of the W3C mailing list test collection reveals that the relevance judgments are generally useful. A Kendall's tau of 0.763 for topical relevance judgments and a tau of 0.776 for pro-con relevance judgments indicate that the rankings of two systems have important differences between the primary and secondary judges but are still significantly correlated. However, the direct inter-assessor agreement scores are relatively low - for the inter-assessor agreements on topical relevance judgments, topic-averaged overlap is 0.29 and overall kappa is 0.42, and for the agreements on pro-con relevance judgments, topic-averaged overlap is 0.275 and overall kappa is 0.3. This indicates that the relevance judgments could be improved. We notice that full relevance judgments across some topic types are more likely to be lower than those across other topic types whereas topical relevance judgements across topic types are more likely to be the same, then we notice that topic type has a significant effect on the agreement of full (but not topical) relevance judgments whereas topic does not; however, the effect of one variable

depends on the level of the other variable.

When looking at the topic categories, we find that some categories have more of a pro/con nature. Intuitively the topics in the B category (method, tip, solution) may not lend to pro/con discussions since methods or solutions (such as ways of fixing a bug) may not generally activate pros or cons. Many of the topics in the C category (i.e., discuss an issue) are vaguely defined, which may lead to difficulties of making pro/con judgments. In other words, at least some of the topics in category B and C may not be appropriate for pro/con relevance judgment and could be better defined for this purpose. This is further validated with our topic and pro/con retrieval system.

The purpose of the inter-assessor agreement analysis across topic types is to design a better test collection. Having a better test collection will give us a better experimental platform for pro/con retrieval. We may improve the W3C mailing list test collection by balancing the topic types. Currently we have half of the topics in category A. Our original expectation was that the topics in the W3C mailing list test collection topics would focus on issues like design rationale, but category F (reason, design rationale) has only 1 topic (Topic 1). Interestingly Topic 1 has high pro/con inter-assessor agreement scores (the agreement overlap is 0.44 and kappa is 0.61) but has no partially relevant documents. We may want to develop more topics in that category.

Fundamentally we need an information needs study for the test collection. So far all the topics have been developed based on assumed user needs and tasks. Since a test collection should be created in a way that models actual information access task, we need to understand what real users want to know from this collection and how real users want to search mailing lists. Our current system could serve as a basis for doing user needs studies and search process research, and understanding user needs would allow us to understand the search processes, which would in turn allow us to design a better system to support the search processes and user needs. Iterations of that cycle will help design a better test collection.

For the DS task of TRECENT 2006, we may find a way to improve the inter-assessor agreement by improving the process by which the test collection is built, such as better defining the topics for pro/con retrieval. Essentially pro/con judgments sometimes are very hard, as experienced by some of the participants of the TRECENT 2005, since determining subjectivity is necessarily a subjective task! Therefore, clear definitions of the topics may help the judges. For instance, giving examples of pro/con statements might be helpful.

5.2 Topic and Pro/Con Retrieval System

To explore use of the test collection, we have designed a topic and pro/con retrieval system in a Rocchio style. Our system achieves a 4.2% improvement of MAP score over our baseline, indicating that the pro/con relevance judgments are useful. However, the major improvements happen to the topics in category A and E, indicating that the topics in A and E are more appropriate or better defined for pro/con judgments than the topics in categories B, C and D.

Much could be done to design a better topic and pro/con classification system. First of all, we could tune various types of weights in the current system. Currently all the pro/con words are assigned a same weight. We could assign higher weights to terms with higher log-odds ratios. We could develop a better weighting scheme for combining the ranked lists generated by topic retrieval and by pro/con retrieval. The round-robin method generates a similar pro/con feature vector for each topic (with minor pairwise difference), hence the pro/con feature vector contributes less variance to the ranking of retrieved documents than the topic terms do. We could also normalize scores across topics in some way, thus mitigating one source of random variation that exists in our present approach.

Second, we have used all the 48 remaining topic for training; however, for a topic in a category, we may apply the remaining topics in this category for training instead. This will allow us to examine the quality of training data category by category.

Third, we may apply other classification methods, such as SVM, Naive Bayes, Maximum Entropy, etc. SVM is considered as good for binary classification tasks and has good performance [6, 32], but it requires lots of training data. We may want to try it out.

Finally, we may develop separate models for detecting pros and cons, and attitudes and opinions. This requires a better understanding of pro/con statements and opinion/attitude expressions. Do we care more about an opinion statement that a piece of software is difficult to use due to a high learning curve or do we care more about an attitude statement that a person still likes the piece of software even though it is difficult to use? If a user's information needs require us to differentiate between them, we would have to address these classification tasks.

6. ACKNOWLEDGMENTS

This work has been supported in part by the Joint Institute for Knowledge Discovery at the University of Maryland. We would like to thank Diego Reforgiato for introducing us to OASIS and Emile Morse and the anonymous reviewers for their helpful comments.

7. REFERENCES

[1] Carmine Cesarano, Bonnie Dorr, Antonio Picariello, et al, 2005. OASYS: An opinion analysis system. *AAAI 2005*.

[2] Nick Craswell, Arjen P. de Vries, and Ian Soboroff, 2005. Overview of the TREC-2005 Enterprise Track. *TREC 2005*, Gaithersburg, Maryland, November 15-18, 2005.

[3] Sanjiv R. Das and Mike Y. Chen , 2006. Yahoo! for Amazon: sentiment extraction from small talk on the web. *8th Asia Pacific Finance Association Annual Conference (2001)*

[4] Kushal Dave, Steve Lawrence, and David M. Pennock, 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. *WWW2003*, May 20-24, 2003, Budapest, Hungary.

[5] Stephen D. Durbin, J. Neal Richter, and Doug Warner, 2003. A system for affective rating of texts. *Proceedings of the KDD Workshop on Operational Text Classification Systems (OTC-3)*, 2003.

[6] George Forman, 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3, 2003, 1289-1305.

[7] Marti Hearst and Ray Larson, 1997. Relevance feedback, SIMS 202 Lecture 25. http://www.sims.berkeley.edu:8000/courses /is202/f98/Lecture25/sld001.htm

[8] David Howell, 2002. Statistical methods for psychology, 5th Ed., Thomson Learning, p167.

[9] Minqing Hu and Bing Liu, 2004. Mining and summarizing customer reviews. *KDD'04*, August 22-25, 2004, Seattle, Washington, USA.

[10] David J. Ittner, David D. Lewis, and David D. Ahn, 1995. Text categorization of low quality images. *Symposium on Document Analysis and Information Retrieval*, Las Vegas, USA, 1995, p301-315.

[11] Alistair Kennedy and Diana Inkpen, 2005. Sentiment classification of movie and product reviews using contextual valence shifters. *FINEXIN 2005*.

[12] M.E. Lesk and G. Salton, 1969. Relevance assessments and retrieval system evaluation. *Information Storage and Retrieval*, 4, 1969, p343-359.

[13] David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li, 2004. RCV1: a new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5, 2004, p361-397.

[14] Jimmy Lin, Eileen Abels, Dina Demner-Fushman, Douglas W. Oard, Philip Wu, and Yejun Wu, 2005. A Menagerie of tracks at Maryland: HARD, Enterpise, QA, and Genomics, oh my! *TREC 2005*, Gaithersburg, Maryland, November 15-18, 2005.

[15] Bing Liu, Minqing Hu, and Junsheng Cheng, 2005. Opinion observer: Analyzing and comparing opinions on the web. *WWW 2005*. May 10-14, 2005, Chiba, Japan.

[16] Satoshi Morinaga, Kenji Yamanishi, et al, 2002. Mining product reputations on the web. *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002)*.

[17] Tetsuya Nasukawa and Jeonghee Yi, 2003. Sentiment analysis: capturing favorability using natural language processing. *K-CAP'03*, October 23-25, 2003, Sanibel Island, Florida, USA.

[18] Scott Nowson, Jon Oberlander, Alastair J. Gill, 2005. Weblogs, Genres, and Individual Differences. *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*, Stresa, Italy, 2005.

[19] Douglas W. Oard, Dagobert Soergel, and David Doermann, et al, 2004. Building an Information Retrieval Test Collection for Spontaneous Conversational Speech. *SIGIR'04*, July 25-29, 2004, Sheffield, South Yorkshire, UK.

[20] Stuart Oskamp, 1991. Attitudes and opinions, 2nd Ed. Englewood Cliffs, New Jersey: Prentice Hall, p12.

[21] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, 2002. Thumbs up? sentiment classification using machine learning techniques. *EMNLP 2002*, 79-86

[22] Bo Pang and Lillian Lee, 2005. See stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *ACL 2005*.

[23] Benjamin Piwowarski and Mounia Lalmas, 2004. Providing consistent and exhausitive relevance assessments for XML retrieval evaluation. *CIKM'04*, November 8-13, 2004, Washington, DC, USA.

[24] J. J. Rocchio, 1971. Relevance feedback information retrieval. *The Smart Retrieval SystemExperiments in Automatic Document Processing*, G. Salton, Ed. Prentice-Hall, Englewood Cliffs, NJ, 313C323.

[25] Mark Sanderson, Justin Zobel, 2005. Evaluation: Information retrieval system evaluation: effort, sensitivity, and reliability. *SIGIR '05*.

[26] Fabrizio Sebastiani, 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, Vol. 34, No. 1, March 2002, p1-47.

[27] Peter Turney, 2002. Thumbs up or thumb down? semantic orientation applied to unsupervised classification of reviews. *ACL-2002*, 417-424.

[28] Ellen M. Voorhees, 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. *SIGIR'98*, Melbourne, Australia.

[29] Ellen M. Voorhees, 2005. Overview of TREC 2005. *TREC 2005*, Gaithersburg, Maryland, November 15-18, 2005.

[30] Theresa Wilson, Janyce Wiebe and Rebecca Hwa, 2004. Just how mad are you? Finding strong and weak opinion clauses. *AAAI-2004*

[31] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann, 2005. Recognizing contextual polarity in phrase-level sentiment analysis. *HLT-EMNLP 2005*.

[32] Yiming Yang and Xin Liu, 1999. A re-examination of text categorization methods. *SIGIR'99*.

[33] Weizhong Zhu, Min Song, and Robert Allen, 2005. TREC 2005 Enterprise Track results from Drexel. *TREC 2005*, Gaithersburg, Maryland, November 15-18, 2005.

APPENDIX

Topic	Type	Ovlp-p	Ovlp-t	Kap-p	Kap-t	Kap-3
3	A0	1	0.75	1	0.85	0.88
6	A0	0.14	0	0.24	0	0.20
8	A0	0.33	0.42	0.47	0.55	0.43
12	A0	0.42	0.32	0.58	0.47	0.43
23	A0	0	0	0	0	0
24	A0	0.18	0.12	0.29	0.14	0.14
28	A0	0.46	0.60	0.62	0.74	0.61
29	A0	0.22	0.29	0.30	0.39	0.31
32	A0	0.21	0.20	0.32	0.29	0.23
35	A0	0	0.04	0	0.05	0.03
38	A0	0.32	0.32	0.38	0.32	0.27
49	A0	0.21	0.40	0.31	0.51	0.34
51	A0	0.17	0.22	0.25	0.27	0.23
52	A0	0.53	0.94	0.69	0.97	0.77
55	A0+B	0.20	0.32	0.28	0.34	0.23
56	A0+D1	0.42	0.49	0.56	0.62	0.51
59	A0+B	0.17	0.10	0.28	0.13	0.10
60	A0	0.08	0.29	0.21	0.42	0.27
A0 Overall		0.26	0.35	0.38	0.37	0.29
5	A1	0.03	0.04	0.05	0.05	0.04
43	A1	0.03	0.11	0.05	0.15	0.08
A1 Overall		0.03	0.08	0	0.02	0
2	A2	0.24	0.14	0.38	0.23	0.20
15	A2	0.71	0.52	0.82	0.68	0.68
36	A2	0.33	0.35	0.37	0.37	0.34
37	A2	0.31	0.42	0.38	0.44	0.36
A2 Overall		0.40	0.36	0.46	0.46	0.40
42	A3	0.28	0.43	0.35	0.47	0.37
48	A3	0.29	0.14	0.44	0.22	0.22
A3 Overall		0.29	0.29	0.38	0.43	0.35
A Overall		0.27	0.32	0.37	0.36	0.29
7	B	0.29	0.35	0.42	0.46	0.39
10	B	0.21	0.25	0.32	0.35	0.25
16	B	0.33	0.30	0.46	0.42	0.40
17	B	0.50	0.80	0.66	0.89	0.72
19	B	0.15	0.20	0.13	0.12	0.09
21	B	0.14	0.43	0.23	0.58	0.38
30	B	0.26	0.12	0.40	0.16	0.13
47	B	0.05	0.29	0.06	0.24	0.08
55	A+B	0.20	0.32	0.28	0.34	0.23
59	A+B	0.17	0.10	0.28	0.13	0.10
B Overall		0.23	0.32	0.26	0.33	0.22
13	C	1	1	1	1	1
14	C	0.50	0.45	0.64	0.58	0.53
25	C	0.30	0.47	0.46	0.62	0.49
27	C	0.10	0.05	0.15	0.05	0.07
31	C	0.35	0.49	0.49	0.62	0.48
34	C	0.20	0.24	0.23	0.16	0.14
45	C	0.11	0.20	0.16	0.29	0.23
46	C	0.73	0.82	0.84	0.90	0.85
C Overall		0.41	0.47	0.38	0.38	0.31
18	D0	0	0.11	0	0.19	0.09
33	D1	0.36	0.35	0.50	0.46	0.40
44	D1	0.21	0.29	0.29	0.32	0.22
56	A0+D1	0.42	0.49	0.56	0.62	0.51
D1 Overall		0.33	0.38	0.50	0.50	0.41
D Overall		0.25	0.31	0.49	0.50	0.41
40	E	0.05	0.12	0.08	0.18	0.11
58	E	0.03	0.55	0	0.42	0.14
E Overall		0.04	0.34	0	0.54	0.24
1	F	0.44	0.32	0.61	0.47	0.35

Table 7: Inter-assessor agreement on 48 topics by topic types in detail.

Note: Notations (such as Ovlp-p, etc.) same as Table 2.