Report on the TREC-5 Experiment: Data Fusion and Collection Fusion

Jacques Savoy, Anne Le Calvé, Dana Vrajitoru

Institut interfacultaire d'informatique
Université de Neuchâtel (Switzerland)
E-mail: Jacques.Savoy@unine.ch
Web page: www.unine.ch/info/

To appear in Proceedings of the TREC'5, D. Harman (Ed.)
NIST Publication 500-238, Gaithersburg (MD), November 1997, pp. 489-502

Summary

This paper describes and evaluates a retrieval model that considers the problem of data fusion and collection fusion as two faces of the same coin. To establish a clear theoretical foundation for combining various sources of evidence provided either by different search schemes (data fusion) or by distributed information services (collection fusion), we have implemented a retrieval model based on the logistic regression methodology.

Participation: Category B, ad-hoc automatic

Introduction

There exist many reasons for considering multiple sources of evidence in information retrieval (Katzer et al., 1982), (Saracevic & Kantor, 1988), (Harman, 1995), and their integration is usually studied in two distinct contexts. Various retrieval strategies or query formulations may operate on the same collection (data fusion problem) (Belkin et al., 1995), (Lee, 1995), subject described in the first part. The second part deals with the collection fusion problem or how distributed information servers may collaborate to answer to a given request (collection fusion problem) (Callan et al., 1995), (Voorhees et al., 1995).

1. Data Fusion Problem

To combine different retrieval schemes (or different query formulations), a retrieval engine might first find the retrieved set associated with each search scheme, and then merge them into a single effective ranked list. To define this underlying merging function, we may consider, for each retrieved record, its rank and / or its retrieval status value. However, the retrieval status values obtained by various weighting schemes may not have a range of possible similar values, leading to a more complex combination situation.

Section 1.1 outlines our test-collection and some evaluations of individual retrieval schemes based on two distinct query constructions. Section 1.2 presents the main design principles of our logistic search model. The last section depicts an evaluation of various suggested schemes and of our approach to data fusion.

1.1. Evaluation of Individual Retrieval Schemes

Before presenting both fusion problems, an outline of our test-collection and an evaluation of some existing retrieval schemes may be useful. For TREC'5, we have considered the WSJ2 corpus (74,520 documents, and 45 queries) as one collection on the one hand, and on the other, as composed of three distinct sub-collections, according to their respective publication year (see Table 1).

Collection	WSJ90	WSJ91	WSJ92	WSJ2
Size	73 Mb	146 Mb	35 Mb	254 Mb
# of documents	21,705	42,652	10,163	74,520
# of topics	38	40	28	45
# relevant doc.	316	602	146	1,064

Table 1: Collection statistics

The queries (#251 to #300) are fully automatically constructed based on the available natural language description. For each sub-collection, we do not use the same number of topics. More precisely, from the WSJ90 collection, the queries {252, 260, 263, 272, 277, 278, 279, 280, 281, 292, 295, 296} do not have any relevant document and are removed from the evaluation. For the WSJ91 corpus, the "null" topics are {253, 260, 262, 263, 267, 276, 278, 279, 281, 296}, and for the WSJ92 collection, the queries {252, 253, 256, 258, 262, 263, 265, 266, 267, 268, 271, 275, 276, 278, 279, 280, 281, 288, 293, 295, 296, 300} are removed for the same reason. For the whole WSJ2 collection, the queries {263, 278, 279, 281, 296} can be ignored.

The evaluation of various vector-processing schemes and the probabilistic OKAPI model is shown in Table 2 within which the OKAPI performance is used as baseline. For this test, each request was constructed based on either the Descriptive section only or on both the Descriptive and Narrative sections (the precise specifications of these search strategies can be found in Appendix 1).

Assuming that a difference of 5% in average precision can be considered as significant, we may conclude that the Narrative section contains important search terms. Thus, the inclusion of this logical section may significantly improve the retrieval effectiveness. As an exception to this rule, the HTN-BNN scheme seems to perform better with short queries than with long requests. From this data, we may also conclude that the OKAPI, LNU-LTC and ATN-NTC search models result in significant enhancement over other vector-processing schemes. Moreover, simple weighting schemes which are based on binary indexing (BNN-BNN) or only on term frequencies (NNN-NNN) must be clearly discarded.

	Precision (% change)	
Collection	WSJ2	WSJ2
	<desc>	<desc>&<narr>
Model	45 queries	45 queries
Individual Retrieval Scheme		
OKAPI - NPN (baseline)	14.06	20.30
LNU - LTC	15.00 (+6.76)	20.47 (+0.84)
ATN - NTC	14.54 (+3.49)	20.48 (+0.89)
LTN - NTC	13.48 (-4.06)	18.58 (-8.47)
LNC - LTC	11.54 (-17.86)	18.51 (-8.82)
LTC - LTC	10.07 (-28.32)	14.65 (-27.83)
ANN - NTC	12.91 (-8.11)	17.04 (-16.06)
ANC - LTC	8.10 (-42.35)	16.39 (-19.26)
HTN - BNN	15.17 (+7.97)	13.35 (-34.24)
LNC - LNC	5.95 (-57.65)	13.20 (-34.98)
ANN - ANN	10.05 (-28.47)	7.80 (-61.58)
NNN - NNN	2.52 (-82.06)	3.68 (-81.87)
BNN - BNN	4.57 (-67.47)	3.45 (-83.00)

Table 2: Evaluation of Individual Retrieval Schemes

1.2. Our Logistic Retrieval Model

From previous research projects, we may conclude that the combination of various retrieval schemes represents a useful strategy for enhancing the retrieval effectiveness, specially when combining multiple queries formulations (Turtle & Croft, 1991), (Shaw & Fox, 1994), (Belkin et al., 1995).

From our point of view, we consider that formulating a request is a difficult task for the users, and asking them to specify two or more queries may render this process more complex. Therefore, we think it is more appropriate to work with a single request formulation. Moreover, most of the previous works make use of heuristics to merge the results of separate search strategies, and only take account of the retrieval status value as an explanatory variable.

To overcome these difficulties, we suggest using the logistic regression (Hosmer & Lemeshow, 1989) as a methodology for combining multiple sources of evidence regarding the relevance of a given document. Of course, this statistical approach has been already applied in other domains such as informetrics (Bookstein et al., 1992) or as a retrieval model (Gey, 1994), (Fuhr & Pfeifer, 1994).

In our approach, we may estimate the probability of a given document D_i's relevance by computing the following formula:

$$\text{Prob}[D_i \text{ is relevant} \mid \mathbf{x}] = \pi(\mathbf{x}) = \frac{1}{1 + e^{-\alpha - \beta \cdot \mathbf{x}}} \qquad (1)$$

within which $\beta \cdot \mathbf{x} = \sum_{j=1}^{r} \beta_{1j} \cdot \text{RANK}_j(D_i) + \beta_{2j} \cdot \text{RSV}'_j(D_i) + \beta_{3j} \cdot \text{VARIA}_j(D_i)$

As shown in Equation 1, our model may take account of the rank, the retrieval status value (without any normalization) and the variation (VARIA) of the retrieval status value compared to the highest value that can be achieved by the corresponding request and search model. For example, within the coordination match model (BNN - BNN), this highest value is defined as the number of search terms.

Based on the WSJ2 corpus and 147 queries (#51 to #250), we have computed the corresponding coefficient values using the SAS package. From the resulting data depicted in Table 3b, we may reach the conclusion that the retrieval schemes

LNU - LTC, LTC - LTC, and LNC - LNC do not have a real impact in our logistic model. These search strategies can thus be ignored (the label "not signif." means that the value of the coefficient can be statistically consider as 0). This conclusion however does not mean that these search models are without any merit, but rather that their influence is already taken into account by the remaining search strategies. For these models, we can also conclude that the rank and either the RSV or the variation are statistically good predictors.

Model	β_{1j} (RANK)	β_{2j} (RSV')	β_{3j} (VARIA)
OKAPI - NPN	not signif.	0.049	not signif.
LNU - LTC	not signif.	not signif.	0.0255
LTN - NTC	-0.00045	0.141	not signif.
LNC - LTC	-0.00033	not signif.	0.0342
LTC - LTC	not signif.	not signif.	0.0197
LNC - LNC	-0.00015	not signif.	not signif.
ATN - NTC	-0.00022	0.195	not signif.
constant α	-6.0871		

Table 3a: Logistic Regression Coefficient Values (Topic = <desc>)

Model	β_{1j} (RANK)	β_{2j} (RSV')	β_{3j} (VARIA)
OKAPI - NPN	-0.00056	not signif.	0.0923
LNU - LTC	not signif.	not signif.	not signif.
LTN - NTC	-0.0004	not signif.	0.00981
LNC - LTC	-0.00051	6.17	not signif.
LTC - LTC	not signif.	not signif.	not signif.
LNC - LNC	not signif.	not signif.	not signif.
ATN - NTC	-0.00064	0.448	not signif.
constant α	-5.8734		

Table 3b: Logistic Regression Coefficient Values (Topic = <desc> & <narr>)

1.3. Evaluation

In evaluating our logistic model, we are also interested to compare its performance with both individual schemes (see Table 2), and with other data fusion strategies. To address this second point, we have implemented a data fusion model derived from the studies of Fox & Shaw (1994) and Lee (1995). After selecting the same individual retrieval strategies (given in Table 4), we first divide the retrieval status value by the maximum of those achieved in the corresponding list (see Equation 2).

$$RSV(D_i) = RSV'(D_i) / Max\{RSV'(D_i)\} \quad (2)$$

in which $RSV'(D_i)$ indicates the retrieval status value obtained by document D_i.

Second, given a set of r retrieval schemes, each producing a normalized retrieval status value $RSV_j(D_i)$, we may combine these multiple sources of evidence according to the following formula:

$$RSV(D_i) = \sum_{j=1}^{r} \alpha_j \cdot RSV_j(D_i) \quad (3)$$

within which the parameters α_j indicate the relative weight associated with each retrieval scheme, and the operator to be applied to combine the retrieval status values. The addition seems to be the best operator (Belkin et al., 1995) and was selected during our evaluation. Moreover, previous reports indicate that an appropriate value for the parameters α_k seems to be a constant (e.g., 1 as defined in our Model 1 in Table 4). However, we may weight the relative importance of each retrieval scheme based on their relative retrieval performance (see Table 2) leading to the definition of our Model 2 in Table 4.

Model	Model 1 α_k	Model 2 α_k
doc. = OKAPI, query = NPN	1	2
doc. = LNU, query = LTC	1	2
doc. = LTN, query = NTC	1	1.5
doc. = LNC, query = LTC	1	1.5
doc. = LTC, query = LTC	1	1
doc. = LNC, query = LNC	1	1
doc. = ATN, query = NTC	1	1.5

Table 4: Parameters Specification

	Precision (% change)	
Collection Model	WSJ2 <desc> 45 queries	WSJ2 <desc>&<narr> 45 queries
Best individual scheme (HTN / ATN)	15.17	20.48
Data fusion Model 1, = SUM	15.15 (-0.13)	22.45 (+9.62)
Data fusion Model 2, = SUM	14.98 (-1.25)	22.58 (+10.25)
Logistic regression $RANK_k(D_i)$, $RSV'_k(D_i)$, $VARIA_k(D_i)$	15.97 (+5.27)	22.72 (+10.94)
Official names	UniNE7	UniNE8

Table 5: Evaluation of Data Fusion Strategies

Compared to the best individual run, our data fusion model presents a significant enhancement. Moreover, the query length seems to play an important role in data fusion Model 1 and 2, leading to the conclusion that such data fusion approach may further improve only long query.

2. Collection Fusion

After selecting the more appropriate sources of information (collection selection problem), a collection fusion strategy must provide a mean of effectively merging multiple independent retrieval results into a single ranked list. Section 2.1 describes related research for resolving the collection fusion problem. Our suggested logistic model is presented in Section 2.2, while the last section depicts an evaluation of some of these strategies.

2.1. Related Works on Collection Fusion

Recent works in this domain have suggested some solutions to the merging of separate answer lists obtained from distributed information services. As a first approach, we might assume that (1) the answer lists obtained from various information servers contain only the ranking of the retrieved items, and (2) that each sub-collection contains roughly the same number of relevant items for each submitted request. In such circumstances, we may interleave the results in a round-robin fashion.

As a second method, we might formulate the hypothesis that each information server applies the same (or very similar) search strategy and that the document score values are directly comparable. Such a strategy, called raw-score merging, produces a final list based on the retrieval status value computed by each sub-collection. However, as demonstrated by Dumais (1993), collection-dependent statistics in document or query weights may vary widely among sub-collections, and therefore, this phenomenon may invalidate the raw-score merging hypothesis.

Finally, Callan et al. (1995) suggest a merging strategy based on the score achieved by both sub-collection and document. Therefore, in this scheme, the sub-collections are ranked according to the probability that they respond appropriately to the current request. This strategy produces a performance similar to a run treating the entire set of documents as a single collection.

2.2. Our Logistic Retrieval Model

In our model, we have analyzed and designed a logistic regression for each information server or sub-collection participating in the final result. To determine the relevance probability for a given document D_i in a given sub-collection, we propose computing the following value:

$$\text{Prob}[D_i \text{ is relevant} \mid \mathbf{x}] = \pi(\mathbf{x}) = \frac{1}{1 + e^{-\beta \cdot \mathbf{x}}} \quad (4)$$

with $\beta \cdot \mathbf{x} = \beta_0 + \beta_1 \cdot \text{RANK}(D_i) + \beta_2 \cdot \text{RSV}'(D_i) + \beta_3 \cdot \text{VARIA}(D_i)$

In this case, the different values $\pi(\mathbf{x})$ may be compared directly with the various sub-collections or search strategies. For the merging procedure, these estimated relevance probabilities define the sort key (or the number and the final position) for each item extracted from each sub-collection. In its actual form, our retrieval scheme does not include a sub-collection selection procedure that, based on the current request, may automatically pick out sub-collections forming part of the final solution. Thus, each query is submitted to all sub-collections and the resulting lists are merged according to the different values of $\pi(\mathbf{x})$.

Model	β_0 (CONSTANT)	β_1 (RANK)	β_2 (RSV')	β_3 (VARIA)
OKAPI - NPN	-5.9763	-0.00317	0.093	0.0839
LNU - LTC	-5.2181	-0.00343	210.9	0.0178
LNC - LTC	-5.9942	-0.00451	22.2	not signif.

Table 6a: Logistic Regression Coefficient Values (Topic = <desc>)

Model	β_0 (CONSTANT)	β_1 (RANK)	β_2 (RSV')	β_3 (VARIA)
OKAPI - NPN	-5.1710	-0.00515	0.028	0.0853
LNU - LTC	-5.1253	-0.00402	220.8	0.0174
LNC - LTC	-5.7964	-0.00668	22.3	0.0

Table 6b: Logistic Regression Coefficient Values (Topic = <desc> & <narr>)

Based on the data in Table 6, one can see that the coefficient values of our logistic model are very similar when comparing the short and long queries.

2.3. Evaluation

As described in Tables 7, our collection fusion problem is particular. As usual, we have divided a test-collection into various sub-collections. However, in this paper, we apply different retrieval schemes to each sub-collection.

Under such circumstances, the round-robin strategy presents relatively interesting performance, while the raw-score merging strategy is clearly ineffective. In fact, all the retrieved documents are extracted form the WSJ90 collection because the OKAPI model retrieval status values are always greater than those of the LNU-LTC or LNC-LTC search schemes (see statistics given in Tables 7). If we normalize the retrieval status value within each sub-collection by dividing them by the maximum RSV of each result list, the retrieval performance is always significantly worse that the round-robin strategy.

Our logistic model presents a significant enhancement over the round-robin scheme when dealing with short queries (Table 7a) and similar performance with long requests (Table 7b).

	Precision (% change)		
Collection	WSJ90	WSJ91	WSJ92
Model	OKAPI - NPN	LNU - LTC	LNC - LTC
	38 queries	40 queries	28 queries
Average precision	17.22	16.29	17.52
# of relevant doc.	316	602	146
# of relevant doc. retrieved	229	400	127
RSV min	2.037	0.002	0.011
RSV max	34.136	0.023	0.309
RSV mean	5.513	0.006	0.0518
RSV standard error	2.433	0.00236	0.0253
Collection Fusion	WSJ2		
	45 queries		
# of relevant doc.	1064		
Round-robin (baseline)	11.38		
Raw-Score Merging	5.89 (-48.24)		
Norm. Raw-Score Merging	9.17 (-19.42)		
Logistic RANK(D_i), RSV'(D_i), VARIA(D_i)	13.35 (+17.31)		
Official name	UniNE0		

Table 7a: Evaluation of Collection Fusion Strategies (Topic = <desc>)

	Precision (% change)		
Collection	WSJ90	WSJ91	WSJ92
Model	OKAPI - NPN	LNU - LTC	LNC - LTC
	38 queries	40 queries	28 queries
Average precision	28.86	19.48	21.45
# of relevant doc.	316	602	146
# of relevant doc. retrieved	249	430	130
RSV min	4.588	0.002	0.017
RSV max	107.889	0.025	0.288
RSV mean	17.937	0.0067	0.05737
RSV standard error	8.926	0.002	0.023
Collection Fusion	WSJ2		
	45 queries		
# of relevant doc.	1064		
Round-robin (baseline)	19.75		
Raw-Score Merging	12.64 (-36.00)		
Norm. Raw-Score Merging	13.18 (-33.27)		
Logistic RANK(D_i), RSV'(D_i), VARIA(D_i)	19.85 (+0.51)		
Official name	UniNE9		

Table 7b: Evaluation of Collection Fusion Strategies (Topic = <desc> & <narr>)

Conclusions and Future Work

This paper describes a unified approach to combining multiple sources of evidence to both the problems of data fusion and collection fusion. To resolve these two distinct questions, we have used the same design principles, algorithm and data structures, showing that the resulting logistic model reveals particularly interesting retrieval effectiveness. Moreover, the evaluation results depicted in this paper demonstrated that the Narrative section of TREC topics has a clear and positive impact on the retrieval effectiveness.

In the near future, we will address the following questions:

a) Data fusion problem: based only on one query formulation, our experience seems to indicate that it is important to consider only two or three retrieval schemes instead of six. Is it always the case and why?

b) Collection fusion problem: when only the rank is available as explanatory variable, how can we use our logistic approach, and does such a retrieval model present a significant enhancement over the round-robin strategy?

c) Are the values of the logistic regression coefficients obtained with one tested collection (WSJ2) valid for another corpus (e.g., SJMN)?

Finally, in this study, we never take known relevance documents (Salton & Buckley, 1990) or pseudo-relevance information into account (Buckley et al., 1995) in order to improve retrieval effectiveness. Although we do not reject this attractive proposition, our objective is to evaluate the effectiveness of the initial search. Relevance feedback can therefore be used after this first search in order to enhance the retrieval performance.

Acknowledgments

The authors would like to thank C. Buckley for giving us the opportunity to use the SMART system, without which this study could not have been conducted. This research was supported by the SNSF (Swiss National Science Foundation) under grant 20-43'217.95.

References

Belkin, N. J., Kantor, P., Fox, E. A. & Shaw, J. A. (1995). Combining the evidence of multiple query representations for information retrieval. *Information Processing & Management*, 31(3), 431-448.

Bookstein, A., O'Neil, E., Dillon, M. & Stephens, D. (1992). Applications of loglinear models for informetric phenomena. *Information Processing & Management*, 28(1), 75-88.

Buckley, C., Singhal, A., Mitra, M. & Salton, G. (1995, November). *New retrieval approaches using SMART*. Proceedings of the TREC'4, Gaithersburg, MD, in press.

Callan, J. P., Lu, Z. & Croft, W. B. (1995, June). *Searching distributed collections with inference networks*. Proceedings of the ACM-SIGIR'95, Seattle, WA, 21-28.

Dumais, S. T. (1993, November). *Latent semantic indexing (LSI) and TREC-2*. Proceedings of TREC'2, Gaithersburg, MD, NIST Publication #500-215, 105-115.

Fuhr, N. & Pfeifer, U. (1994). Probabilistic information retrieval as a combination of abstraction, inductive learning, and probabilistic assumptions. *ACM Transactions on Information Systems*, 12(1), 92-115.

Gey, F. C. (1994, July). *Inferring probability of relevance using the method of logistic regression*. Proceedings of the ACM-SIGIR'94, Dublin, Ireland, 222-231.

Harman, D. (1995). Overview of the second text retrieval conference (TREC-2). *Information Processing & Management*, 31(3), 271-289.

Harman, D. (1986). *An experimental study of factors important in document ranking*. Proceedings of the ACM-SIGIR'86, Pisa, Italy.

Hosmer, D. & Lemeshow, S. (1989). *Applied logistic regression*. New-York, NY: John Wiley & Sons.

Katzer, J., McGill, M. J., Tessier, J. A., Frakes, W. & DasGupta, P. (1982). A study of the overlap among document representations. *Information Technology: Research & Development*, 2, 261-274.

Lee, J. H. (1995, July). *Combining multiple evidence from different properties of weighting schemes*. Proceedings of the ACM-SIGIR'95, Seattle, WA, 180-188.

Salton, G. & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4), 288-297.

Saracevic, T. & Kantor, P. (1988). A study of information seeking and retrieving. III. Searchers, searches, overlap. *Journal of the American Society for Information Science*, 39(3), 197-216.

Shaw, J. A. & Fox, E. A. (1994, November). *Combination of multiple searches*. Proceedings of the TREC'3, Gaithersburg, MD, NIST Publication #500-225, 105-108.

Turtle, H. & Croft, W. B. (1991). Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3), 187-222.

Voorhees, E. M., Gupta, N. K. & Johnson-Laird, B. (1995, July). *Learning collection fusion strategies*. Proceedings of the ACM-SIGIR'95, Seattle, WA, 172-179.

Appendix 1: Weighting Schemes

In this paper, the indexing procedure done by the SMART system is fully automatic and based on a single term only. The representation of each topic is based on the content of its Descriptive (<desc>) section or its Descriptive and Narrative (<narr>) sections. For each document, the Text (<text>) section as well as the Subtitle (<ST>), Headline (<HL>), and Summary (<LP>) sections were used to build the document surrogate. All other subsections were removed, and, in particular, the title and the concept section of each topic (see Table A.1).

Collection	Section
WSJ2	<desc>, <text>, <st>, <hl>, <lp>
Query	<desc> or <desc> & <narr>

Table A.1: Selected Sections Used to Represent Documents and Queries

To assign an indexing weight w_{ij} reflecting the importance of each single-term T_j, j = 1, 2, ..., t, in a document D_i, we may use one of the equations shown in Table A.2. In this table, tf_{ij} depicts the frequency of the term T_j in the document D_i (or in the request), n represents the number of documents D_i in the collection, df_j the number of documents in which T_j occurs, and idf_j the inverse document frequency ($\log [n / df_j]$). Moreover, the document length of D_i (the number of indexing terms) is noted by nt_i, and $mean(nt_\cdot)$ indicates the collection mean. The constant c is fixed to 0.2 and C is computed as $0.5 + 1.5 \cdot [nt_i / mean(nt_\cdot)]$. Finally, the computation of the retrieval status value is based on the inner product.

BNN	$w_{ij} = 1$	NNN	$w_{ij} = tf_{ij}$
ANN	$w_{ij} = 0.5 + 0.5 \cdot \frac{tf_{ij}}{\max tf_{i.}}$	ATN	$w_{ij} = \left[0.5 + 0.5 \cdot \frac{tf_{ij}}{\max tf_{i.}}\right] \cdot idf_j$
NPN	$w_{ij} = tf_{ij} \cdot \log \left[\frac{n - df_j}{df_j}\right]$	LTN	$w_{ij} = [\log(tf_{ij})+1] \cdot idf_j$
HTN	$w_{ij} = \frac{\log(tf_{ij}+1) \cdot idf_j}{\log(nt_i)}$	OKAPI	$w_{ij} = \sum_{k=1}^{t} \frac{2 \cdot tf_{ik}}{C + tf_{ik}}$
LNC	$w_{ij} = \frac{\log(tf_{ij})+1}{\sqrt{\sum_{k=1}^{t} (\log(tf_{ik})+1)^2}}$	NTC	$w_{ij} = \frac{tf_{ij} \cdot idf_j}{\sqrt{\sum_{k=1}^{t} (tf_{ik} \cdot idf_k)^2}}$

ANC	$w_{ij} = \frac{0.5 + 0.5 \cdot \frac{tf_{ij}}{\max tf_{i.}}}{\sqrt{\sum_{k=1}^{t} \left[0.5 + 0.5 \cdot \frac{tf_{ik}}{\max tf_{i.}}\right]^2}}$
LTC	$w_{ij} = \frac{[\log(tf_{ij})+1] \cdot idf_j}{\sqrt{\sum_{k=1}^{t} ([\log(tf_{ik})+1] \cdot idf_k)^2}}$
LNU	$w_{ij} = \frac{\frac{1+\log(tf_{ij})}{1+\log(mean\ (tf_{i.}))}}{(1 - c) \cdot mean(nt_{.}) + c \cdot nt_i}$

Table A.2: Weighting Schemes