B. Mitschang et al. (Hrsg.): BTW 2017 – Workshopband,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2017 301

Comparing Relevance Feedback Techniques on German News Articles

Julia Romberg[1]

Abstract: We draw a comparison on the behavior of several relevance feedback techniques on a corpus of German news articles. In contrast to the standard application of relevance feedback, no explicit user query is given and the main goal is to recognize a user's preferences and interests in the examined data collection. The compared techniques are based on vector space models and probabilistic models. The results show that the performance is category-dependent on our data and that overall the vector space approach *Ide* performs best.

Keywords: Relevance Feedback, Text Mining, Filtering Systems

1 Introduction

Over the last decades personalization has become of major interest and it continually gains on popularity. Especially in information systems, involving a large amount of data, a user benefits from the presentation of only a subset of data, which is customized to the user's particular information needs or general interests.

In case of a target-oriented information need, information retrieval systems serve to satisfy this as follows: A user formulates a query that fits his or her specific need for information. The information system then returns the most appropriate documents regarding the query using some algorithm. These documents are all assumed to be relevant for the given information demand. However, this may not always be the case in real-world applications as the envisioned request usually does not completely match the phrased query. One example is the search for documents related to the animal crane. A user, with a clear objective in mind, may therefore formulate the simple query "crane". Using Google search the first results include documents that are about the animal, about the machine type called crane, or even about a paper manufacture carrying "Crane" in its name. As can be seen from this example, a semantic gap between imagination and the query can arise. To nevertheless ensure that the user's information needs will be satisfied, the formulated query can be adjusted using relevance feedback [RS65]: Following the information retrieval step, the user rates the best matching documents as relevant or non-relevant. Involving this assessment, the initial query can be modified in order to reduce the semantic gap and hence better results can be achieved.

Whereas traditional information retrieval systems serve a target-oriented information need, the more general recognition of user interests is mostly done by information filtering

[1] Heinrich Heine University Düsseldorf, Institute of Computer Science, Universitätsstraße 1, 40225 Düsseldorf, romberg@cs.uni-duesseldorf.de

systems. Information filtering systems aim to help finding user-preference-based documents by filtering unwanted information out of an overload of information. If we treat filtering as a text classification task with two classes (relevant and non-relevant), techniques such as neural networks, support vector machines, decision trees, or logistic regression, tend to perform rather well once some basis knowledge of the user's preferences is given [Zh09]. In order to collect initial training data and to handle the so-called cold start problem, target-oriented retrieval techniques may be used.

The purpose of our work is to compare different retrieval techniques that involve relevance feedback with regard to their behavior on the cold start problem and on a corpus of German news articles. The considered models are on the one hand vector space models and on the other hand models that originate in probabilistic theory. The remainder of this paper is structured as follows: In the next section we discuss important papers for ad-hoc retrieval. We then describe the data corpus, the relevance feedback techniques on which we focus, and the query formulation. Subsequently the techniques are evaluated and compared. Finally a conclusion is drawn and further work is outlined.

2 Related Work

A lot of research has been done in the field of ad-hoc retrieval and on retrieval models that involve relevance feedback. The best known approaches seem to be query modification in vector space models and improved relevance guesses in probabilistic information retrieval systems.

The first relevance feedback approach, introduced in [RS65], relates to the vector space model and gives a formula to spatial shift a vector that represents the user query towards points that represent relevant documents. In [Id71] this idea is taken up on with changed query modification formulas.

Probabilistic models estimate the relevance of documents. An overview over models such as the basis model [MK60], the binary independence model [RJ76, Ri79] and the Okapi BM25 model [RW94] is given by [JWR00a, JWR00b]. In general, relevance estimation is done by making use of a subset of previously rated documents. This intuitively leads to the application of relevance feedback. Further models that also originate in probability theory, for example Language models [HR01] and Bayesian networks [TC89], have also been applied in connection with relevance feedback.

The approaches have mostly been tested on TREC collections[3] that were released for tasks of the Text REtrieval Conference. We, however, want to evaluate on a different set of data: German news articles taken from online resources. We want to observe how the compared techniques behave on this specific data source, whether their performance is topic-dependent, and which one performs best.

[3] http://trec.nist.gov

3 Data and Relevance Feedback Methods

In this section, first, the data corpus is presented. We then illustrate the data representation we use for the comparison. Afterwards, an overview of the used relevance feedback techniques is given. Finally, we focus on the challenge of lacking an initial query.

3.1 Corpus

One basic requirement for a good comparability of different relevance feedback techniques on a quantity of data is a high linguistic quality as well as good content-related quality. This is essential as the corpus needs to be transferred in a comparable representation using a natural language processing pipeline, whereby errors should be minimized, and also as we evaluate on test persons. The German online platform *ZEIT ONLINE*[4] fulfills these conditions. Furthermore, it provides an API to get access to the meta data of all archived articles.

Our data corpus consists of 2500 German news articles which were crawled from *ZEIT ONLINE*. They were divided up in the categories *politics*, *society*, *economy*, *culture*, and *sports*, with a proportion of 500 articles each. This size was selected to fit on the one hand the need for a sufficiently large corpus to get significant results on the evaluation runs and to simultaneously reduce the impact older articles have on relevance ratings. News actuality often correlates with relevance so that the given relevance feedback could be biased by including articles that have no relation to current events.

3.2 Data Representation

In order to run our evaluation, we first need to transform the collection of documents in an applicable format with regard to the chosen relevance feedback methods. The most general approach is the use of a bag-of-words: Given an index term vocabulary $V = \{w_1, ..., w_n\}$, a document is expressed by a n-dimensional vector in which every entry is the term frequency of w_i in the regarded document. Important contents of news articles, particularly online, use to be described by key words. These key words are mostly nouns and named entities, such as person names, locations, and companies. We therefore decide to only include nouns and named entities in V. This also leeds to a remarkable dimension reduction. While in the bag-of-words model words are seen as isolated units, topic models attempt to incorporate relations between words by uncovering underlying semantic structures of the collection. One of the best known topic models is Latent Dirichlet Allocation [BNJ03]. The main idea here is that a collection covers k topics. Thereby a topic is defined as a distribution over a fixed index term vocabulary and each document belongs, to a certain proportion, to each of the k topics. A document is represented by a k-dimensional vector of length 1, which consists of the topic membership proportion.

[4] http://www.zeit.de

3.3 Vector space models

Vector space models are based on the assumption that documents and querys can be represented as vectors located in a vector space. A document's relevance concerning a given query is defined by the similiarity between the two representing vectors. We used the cosine similarity, which is the most commonly used similarity measure in this context: The smaller the angle between two vectors is, the more alike they are.

The first approach we use is *Rocchio*'s formula [RS65]. On the basis of subsets with known relevance judgments, the spatial position of an initial query vector $\vec{q}$ is improved by moving the original query vector $\vec{q}$ towards the centroid of the relevant documents D_r and away from the centroid of the non-relevant documents D_{nr}:

$$\vec{q}_m = \alpha \cdot \vec{q} + \beta \cdot \frac{1}{|D_r|} \cdot \sum_{\forall \vec{d}_j \in D_r} \vec{d}_j - \gamma \cdot \frac{1}{|D_{nr}|} \cdot \sum_{\forall \vec{d}_j \in D_{nr}} \vec{d}_j \tag{1}$$

$\overrightarrow{q_m}$ denotes the modified query vector which is then used to calculate the similarity to the document vectors in order to determine a more appropriate set of relevant documents. The weights $\alpha, \beta, \gamma \in \mathbb{R}_0^+$ control the impact of the single components.

One further vector space approach is *Ide* [Id71]. It removes the centroid normalization from formula 1:

$$\vec{q}_m = \alpha \cdot \vec{q} + \beta \cdot \sum_{\forall \vec{d}_j \in D_r} \vec{d}_j - \gamma \cdot \sum_{\forall \vec{d}_j \in D_{nr}} \vec{d}_j \tag{2}$$

3.4 Probabilistic models

In contrast to vector space models, classic probabilistic models aim to estimate a document's relevance for a given query. For this, a term weight $\in [0,1]$ is assigned to every query term that expresses the probability of obtaining a relevant document by this term.

We used the *Binary Independence Model* and the *Okapi BM25 Model* with a smoothed term weighting proposed in [RJ76]:

$$\widetilde{w_i} = \frac{|D_{r_i}| + 0.5}{|D_r| - |D_{r_i}| + 0.5} \cdot \frac{|D| - |D_i| - |D_r| + |D_{r_i}| + 0.5}{|D_i| - |D_{r_i}| + 0.5},$$

where D_{r_i} denotes the relevant documents which contain the ith term out of all relevant documents D_r and D_i denotes the number of documents out of all documents $D = D_r \cup D_{nr}$ that contain this term.

The overall estimated relevance for a document d and for a given query $q = (w_1, w_2, ..., w_t)$ in the Binary Independence Model is, under the assumption of term independence, provided by the logarithm of the product over all weights $\widetilde{w_i}$ of the query terms that are contained in d:

$$relevance\_score(d,q) = \log \prod_{i=1}^{t} \widetilde{w_i} \cdot \mathbb{1}_{i \in d} = \sum_{1=i}^{t} \log(\widetilde{w_i}) \cdot \mathbb{1}_{i \in d} \tag{3}$$

Additionally, in the Okapi BM25 Model term frequency and document length are taken into account by

$$relevance\_score(d,q) = \sum_{i=1}^{t} \log(\widetilde{w}_i) \cdot \frac{\mathrm{tf}(w_i,d)\cdot(k_1+1)}{k_1\cdot((1-b)+b\cdot\frac{\mathrm{dl}}{\mathrm{avdl}})+\mathrm{tf}(w_i,d)} \tag{4}$$

The parameter k_1 controls the influence of w_i's term frequency $\mathrm{tf}(w_i,d)$ in d. The impact of the document length dl in proportion to the average document length avdl in the collection is given by $b \in [0,1]$.

Intuitively, a good way to reflect the user's information needs would be a query which consists of terms that occur frequently in relevant documents. This idea is the basis for a different approach to probabilistic models named language models. The probability that a document is relevant to a given query can, according to that, be estimated by

$$P(d|q=(w_1,w_2,...,w_t)) = P(d)\cdot P(q|d) = P(d)\cdot\prod_{i=1}^{t}((1-\lambda_i)P(w_i)+\lambda_i P(w_i|d)) \tag{5}$$

A permissible way to estimate the probabilities is [HR01]:

$$P(w_i) = \frac{\mathrm{df}(w_i)}{\sum_{j=1}^{t}\mathrm{df}(w_j)}, \quad P(w_i|d) = \frac{\mathrm{tf}(w_i,d)}{\sum_{j=1}^{t}\mathrm{tf}(w_j,d)}, \quad P(d) = \frac{\sum_{j=1}^{t}\mathrm{tf}(w_j,d)}{\sum_{j=1}^{t}\sum_{k=1}^{|D|}\mathrm{tf}(w_j,d_k)},$$

where $\mathrm{df}(w_i)$ denotes the document frequency of the query term w_i.

The parameter $\lambda_i \in [0,1]$ mirrors the importance of the ith term in q. In order to find a good value for λ_i we use an expectation-maximization algorithm as suggested in [HR01]. Expectation-maximization is a statistical method to estimate unknown parameters on incomplete data: By alternating an expectation step and a maximization step, conclusions on the probability distribution in the complete data set can be drawn. The expectation step is defined as

$$r_i' = \sum_{j=1}^{|D_r|} \frac{\lambda_i^{(p)} P(w_i|d_j)}{(1-\lambda_i^{(p)})P(w_i)+\lambda_i^{(p)}P(w_i|d_j)}$$

and the maximization step is defined as

$$\lambda_i^{(p+1)} = \frac{r_i'}{|D_r|}.$$

3.5 Query Formulation

The precondition for relevance feedback techniques to work is the existence of an initial query. In our case, there is no user formulated query to start with. To overcome this problem,

we initially define $\vec{q}$ in vector space models as zero vector, equivalent to the document representation with n dimensions in the bag-of-words format and k dimensions in the LDA format. After a first training phase, the vector is then modified into a more meaningful vector. On the other hand in probabilistic models we initially admit every index term as possible query term.

4 Evaluation

In order to run the comparison of the relevance feedback methods, the experimental setup is introduced at first. After that, different evaluation measures are discussed, followed by the evaluation results which are described and then discussed.

4.1 Experimental Setup

Table 1 gives an overview over the relevance feedback techniques that are compared. As most of the techniques' performances depend on the parameter selection, we tried different settings that were current in the literature. Furthermore the vector space model techniques were evaluated on both data representations: bag-of-words and LDA. Beside the comparison between techniques, we also wanted to observe if the techniques behave differently on various news article categories. We exemplarily took the categories *culture*, *economics*, *politics*, *society* and *sports*.

The ideal experimental setup would imply for every test person to evaluate any technique on any category and over a period of time, i.e. to run several feedback iterations. Unfortunately

Technique	Parameter Setting	BOW	LDA	Abbreviation
Rocchio (formula 1)	$\alpha = 1, \beta = 0.75, \gamma = 0.25$	✓		R
Rocchio (formula 1)	$\alpha = 1, \beta = 0.75, \gamma = 0.25$		✓	LR
Rocchio (formula 1)	$\alpha = 1, \beta = 1, \gamma = 0$	✓		R2
Rocchio (formula 1)	$\alpha = 1, \beta = 1, \gamma = 0$		✓	LR2
Ide (formula 2)	$\alpha = 1, \beta = 1, \gamma = 1$	✓		Ide
Ide (formula 2)	$\alpha = 1, \beta = 1, \gamma = 1$		✓	LI
Ide (formula 2)	$\alpha = 1, \beta = 1, \gamma = 0$	✓		I2
Ide (formula 2)	$\alpha = 1, \beta = 1, \gamma = 0$		✓	LI2
Binary Independence Model (formula 3)		✓		BIM
Okapi BM25 (formula 4)	$b = 0.75, k_1 = 1.2$	✓		B25
Okapi BM25 (formula 4)	$b = 0.75, k_1 = 1.6$	✓		252
Okapi BM25 (formula 4)	$b = 0.75, k_1 = 2.0$	✓		253
Language Model (formula 5)	EM-algorithm	✓		LM

Tab. 1: List of evaluated techniques and abbreviation names

this was not practically applicable with the test persons, which were not rewarded. The main goal of the comparison is the comparison between the techniques. For this reason we decided for every test person to evaluate all techniques but not all categories. We had 29 test persons which were randomly assigned to the five categories having five persons on culture and six persons on each other category. Instead of comparing the techniques over several iterations we chose a train-and-test setup with a bigger training phase: For the training the test persons had to give relevance feedback on 100 randomly chosen news articles. For the testing the test persons had to give relevance feedback on the top 25 results calculated by every technique-setting pair.

To observe the adaption on new data, in the test phase news articles that had already been shown and rated during the training were not shown again. The test persons were furthermore explicitly requested to rate the relevance of an article independent from up-to-dateness.

4.2 Evaluation Measures

The evaluation of relevance feedback is challenging as it is quite subjective. An appropriate way to evaluate a classification task is to calculate the *precision*:

$$\text{precision} = \frac{|\text{relevant} \cap \text{retrieved}|}{|\text{retrieved}|}$$

Precision is a set-based measure. It does not take the degree of similarity of a document and a query into account. However, in our case retrieval is done by determing the k best results, which implies a ranking. Therefore a ranking-based measure could be more informative. A widely used one is *MAP*:

$$MAP = \frac{1}{N} \cdot \sum_{i=1}^{N} \left(\frac{1}{|D_{r_i}|} \cdot \sum_{k=1}^{n} \left(precision@k \cdot \mathbb{1}_{\{k \in D_{r_i}\}} \right) \right)$$

MAP is the mean average precision of a given ranking over N instances. The average precision *AP* is calculated from the precision at any position $k \in \{1,...,n\}$ that was rated as relevant for the instance i, denoted here by the indicator function $\mathbb{1}_{\{k \in D_{r_i}\}}$. In our case, the instances i represent the test persons.

For every technique precision and *AP* are calculated for all test persons and then averaged to have a comparative value. Somewhat problematic thereby is the comparability of different test persons. Every test person has a different affinity towards the assigned topic which results in differences in distribution and proportion of relevant documents. To bypass this circumstance the application of the *argmax* may be the best evaluation measure:

$$\underset{x \in T}{\text{argmax}}\, f(x)$$

T is the set of techniques that are evaluated. The function value $f(x)$ is the total number of test persons for which technique x performed best, either by using precision or *AP*.

technique	mean precision (variance)	*MAP* (variance)
R	56.1 (4.9)	42.1 (6.7)
R2	54.1 (5.7)	39.1 (6.4)
Ide	61.4 (6.2)	49.0 (9.6)
I2	59.3 (5.4)	45.8 (7.3)
BIM	60.0 (5.1)	45.7 (7.3)
B25	61.4 (5.0)	47.2 (7.2)
252	61.2 (5.1)	47.1 (7.4)
253	61.1 (5.3)	47.3 (7.7)
LM	59.0 (4.6)	42.5 (5.8)
LR	55.6 (4.9)	38.2 (4.9)
LR2	53.5 (4.6)	35.7 (4.7)
LI	58.5 (6.1)	42.8 (6.9)
LI2	57.1 (5.5)	41.7 (5.7)

Tab. 2: mean precision and *MAP* without compliance of category

	R	R2	Ide	I2	BIM	B25	252	253	LM	LR	LR2	LI	LI2
culture	0/0	0/0	2/1	1/1	1/1	1/0	2/1	3/1	0/0	0/0	0/0	1/0	0/0
economics	2/1	0/0	2/0	1/0	1/1	2/0	2/1	1/0	1/1	2/1	2/1	2/1	2/1
politics	2/2	4/3	1/2	2/1	0/0	1/1	0/0	0/0	1/0	0/0	0/0	1/0	0/0
society	1/1	0/0	4/4	0/0	1/1	1/1	1/1	1/1	1/1	0/0	0/0	1/1	1/1
sports	0/1	1/0	0/0	0/0	1/1	0/0	0/0	0/0	0/0	1/1	2/0	3/2	1/1
all	5/5	5/3	9/7	4/2	4/4	5/2	5/3	5/2	3/2	3/2	4/1	8/4	4/3

Tab. 3: *argmax* of precision / *argmax* of *AP*

4.3 Results

We now want to present and discuss the evaluation results. Table 2 shows the mean precision and the *MAP* for every technique without compliance of the category that the test persons were split into. In addition, the variance within the techniques is stated. The best results for mean precision are reached by Ide and by B25 with 61.4, whereby B25 shows a slightly lower variance. The *MAP* shows similar results with Ide being the best, achieving a value of 49.0. B25, however, according to this measure, performs worse (47.2). If we only focus on the techniques performed on the LDA-representation, both measures determine LI as the best. Table 3 shows the results evaluated by means of the argmax-function. The first five rows stand for the categories and the columns show the techniques. The row „all“ gives a category-independent overview of the number of test persons that each technique performed best for. For mean precision, Ide and LI clearly outperform the other techniques with a total of 9 or 8. *MAP* shows the same tendency to a lesser extent.

Mean precision and *MAP* values for the category-dependent consideration of the results are shown in Table 4 and Table 5. The columns represent the different approaches while the

	R	R2	Ide	I2	BIM	B25	252	253	LM	LR	LR2	LI	LI2
culture	53.6	50.4	68.0	66.4	59.2	62.4	64.8	67.2	57.6	54.4	49.6	65.5	57.6
economics	60.7	62.0	62.0	61.3	68.7	70.0	68.7	66.7	62.0	65.3	62.0	64.7	67.3
politics	63.3	64.7	59.3	60.7	56.0	58.7	58.7	58.7	61.3	46.7	46.0	53.3	48.7
society	60.7	56.7	76.7	66.7	64.0	70.7	71.3	71.3	67.3	64.0	60.0	62.8	65.3
sports	42.0	36.0	42.0	42.7	52.0	45.3	43.3	42.7	46.7	47.3	49.3	47.3	46.7

Tab. 4: mean precision

	R	R2	Ide	I2	BIM	B25	252	253	LM	LR	LR2	LI	LI2
culture	38.1	33.4	57.8	56.4	46.6	51.4	53.7	55.5	37.6	33.3	31.3	45.3	42.7
economics	46.8	45.6	47.6	43.7	56.6	55.8	54.6	53.1	44.7	48.4	44.1	51.4	49.5
politics	50.6	51.5	46.8	48.5	43.7	45.0	44.6	44.8	43.8	30.1	30.5	34.5	31.0
society	46.4	39.8	67.3	55.2	48.0	54.6	54.7	55.0	55.4	48.3	42.9	46.2	51.0
sports	27.9	24.2	26.8	27.2	33.8	29.9	29.1	29.3	30.1	30.3	29.1	37.1	34.4

Tab. 5: *MAP*

rows represent the five categories. For the categories culture and society, Ide leeds to the best values, having mean precision= 68.0, $MAP = 57.8$ for culture and mean precision= 76.7, $MAP = 67.3$ for society. B25 obtains a mean precision of 70.0 and BIM a *MAP* of 56.6 as best for economics. In politics, R2 holds the leading position, with mean precision= 64.7 and $MAP = 51.5$. Finally, for sports BIM yields the best mean precision (52.0) and LI yields the best *MAP* (37.1). Focusing on the worst results, for any category and for both measures surprisingly Rocchio's algorithm with the proposed parameter settings performs the worst. Table 3 shows the evaluation received by the argmax. In contrast to the category-independent consideration, no clear tendency can be observed, which underlines the difference in the application of the techniques between the categories.

In summary, a dependency of category and best performing technique is apparent on our test data and test persons. Depending on the measurement, Ide's approach and the Binary Independence Model as well as the Okapi BM25 Model perform the best, whereby Ide seems to slightly lead. Rocchio's algorithm produced the worst results according to the three measures we used, although, in a few cases, it also performed well.

5 Conclusion and Future Work

In this work different approaches for the application of relevance feedback and ad-hoc retrieval techniques have been compared on a data corpus including German news articles. Thereby the focus was on the performance of the single techniques, category-dependent as well as category-independent. We showed that the techniques are category-dependent on our data representation and that Ide's formula (Formula 2) slightly outperforms the other techniques by means of the applied measures. Surprisingly Rocchio's formula (Formula 1)

did not achieve similar results to Ide's, although this is frequently stated in the literature. Nevertheless we have to admit that the determination of a appropriate measure for relevance feedback is challenging.

In future work, we want to extend the evaluation. Further parameter settings, a larger number of test persons and other data representations could lead to more meaningful results. We also plan to analyze a larger collection and to test the results we achieved on significance. Furthermore, an evaluation over all techniques and all categories for one test person, and over several iterations, would be interesting. This would also enable to evaluate approaches that refer to rank-based result sets, for example *Ide Dec-Hi* [Id71].

References

[BNJ03] Blei, David M; Ng, Andrew Y; Jordan, Michael I: Latent Dirichlet Allocation. Journal of machine Learning research, 3(Jan), 2003.

[HR01] Hiemstra, Djoerd; Robertson, Stephen E.: Relevance Feedback for Best Match Term Weighting Algorithms in Information Retrieval. In: DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries. 2001.

[Id71] Ide, Eleanor: New Experiments in Relevance Feedback. The SMART retrieval system, 1971.

[JWR00a] Jones, K Sparck; Walker, Steve; Robertson, Stephen E.: A Probabilistic Model of Information Retrieval: Development and Comparative Experiments: Part 1. Information Processing & Management, 36(6), 2000.

[JWR00b] Jones, K Sparck; Walker, Steve; Robertson, Stephen E.: A Probabilistic Model of Information Retrieval: Development and Comparative Experiments: Part 2. Information Processing & Management, 36(6), 2000.

[MK60] Maron, M. E.; Kuhns, J. L.: On Relevance, Probabilistic Indexing and Information Retrieval. J. ACM, 7(3), 1960.

[Ri79] Rijsbergen, C. J. Van: Information Retrieval. 2nd edition, 1979.

[RJ76] Robertson, S. E.; Jones, K. Sparck: Relevance Weighting of Search Terms. Journal of the American Society for Information Science, 27(3), 1976.

[RS65] Rocchio, JJ; Salton, G: Information Search Optimization and Interactive Retrieval Techniques. In: Proceedings of the November 30–December 1, 1965, fall joint computer conference, part I. ACM, 1965.

[RW94] Robertson, Stephen E; Walker, Steve: Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In: Proceedings of the 17th annual international ACM SIGIR conference on Research and Development in Information Retrieval. Springer-Verlag New York, Inc., 1994.

[TC89] Turtle, Howard; Croft, W Bruce: Inference Networks for Document Retrieval. In: Proceedings of the 13th annual international ACM SIGIR conference on Research and development in information retrieval. 1989.

[Zh09] Zhang, Yi: Text Mining: Classification, Clustering, and Applications. RC Press, chapter Adaptive Information Filtering, 2009.