Information filtering using probabilistic model

Hidekazu Yanagimoto
Department of Computer and Systems Sciences
Osaka Prefecture University
1-1, Gakuencho, Sakai
Osaka 599-8531, Japan

Sigeru Omatu
Department of Computer and Systems Sciences
Osaka Prefecture University
1-1, Gakuencho, Sakai
Osaka, 599-8531, Japan

({hidekazu, omatu}@cs.osakafu-u.ac.jp)

Abstract

We propose an information filtering system based on a probablistic model. We have an assumption that a document is generated according to a probability distribution and regard a document as a sample drawn according to the distribution. In this paper, we adopt multinomial distribution and represent a document as probabilitiy which has random values as words in the document. When an information filtering system selects information, it uses a similarity between a user profile and a document. Since our proposed sytem is constructed under the probalistic model, the similarity is defined using Kullback-Leibler divergence. To create the user profile, we must optimize the Kullback-Leibler divergence. Since Kullback-Leibler divergence is non-linear function, we use a genetic algorithm to optimize it. We carry out experiments and confirm effectiveness of the propsed method.

Keyword: Information Filtering, Genetic Algorithm, Kullback Leibler Divergence

1 Introduction

Users can have access to vast store of information through the Internet recently, as the Internet becomes widespread. When they search information they desire with search engines, they get a lot of search results and hardly find ones they desire truly. To cope with this situation, many researchers have paid attention to developing information filtering systems. The information filtering systems automatically select information depending on user's interests (user profile). To select information correctly, the systems must create the user profile which represents user's interests exactly.

Many researchers pay attention to studying a probablistic model of documents. This model is different from vector space model (VSM) which is used in many information retirieval systems. In the VSM, documents are described as vectors using term frequency and document frequency. On the other hand, doucments are described as a probability distribution in the probabilistic model. This causes improving information filtering systems.

In this paper, we propose an information filtering system based on a probablistic model. A similarity between a user profile and a document is defined using Kullback-Leibler divergence (KL divergence). The KL divergence is a measure which represents a similarity between two probability distributions. Since the KL divergence is non-linear, a GA explores the user profile. Since our aim is to improve an information filtering system, we discuss effectiveness of a proposed mnethod in the viewpoint of selection ability.

2 Previous work

An information filtering system selects information depending on user's interests. Hence, it is important for the system to keep user's interests (user profile) exactly. To make the user profile, information retrieval techniques, pattern recognition and machine learning are used.

To deal with documents on computers, VSM and relevance feedback are reported by Salton et al[1]. In the VSM all documents are represented as vectors and a distance between documents is defined using an inner produce between the documents. To clear up user's interest in the VSM, relevance feedback is used.

A probabilistic approache is reported to enrich a document vector[3]. In this approache, we assume documents are generated from an unknow distribution. Hence, our aim is to identify the unknown disribution. In the probabilistic model, the distribution is regarded

as a binomial distribution or a multinomial distribution.

A GA is an algorithm which simulates biological evolution and is proposed by Holland[2]. The GA is often applied to combinational optimization problems and can optimize a non-linear function.

3 Methodology

In this section we explain key concepts of a proposed mnethod. They are a probablistic model, a similarity using Kullback-Leibler divergence and a GA.

3.1 Probabilistic Model

In a probabilistic model it is assumed that that a document is generated from an unknown probability distribution. Hence, a main aim is to identify the unknown distribution from sample documents. In probabilistic information retrieval, the unknow distribution is restrictive. For example, we assume that the distribution is a binomial distribution or a multinomial distribution. In this paper, we adopt the assumption that the distirbution is a multinomial distribution.

The multinomial distribution is represented below.

$$Pr(X_1 = n_1, \cdots, X_k = n_k) = \frac{n!}{\prod_{i=1}^{k} n_i!} \prod_{i=1}^{k} p_i^{n_i} \quad (1)$$

where X_i is a random variable, p_i is a probability of the ith random variable X_i, n_i is a frequency of X_i and n is a sum of n_i. Since a document is generated from the multinomial distribution in this paper, we estimate p_i in Equation(1) using a maximum likelihood estimate method. THe maximum likelihood estimate is calculated below.

$$p_i = \frac{n_i}{n} \quad (2)$$

We estimate one multinomial distribution from one document and consider the distribution represents the document. We define a user profile as a probability distribution, too . However, we do not assume that the user profile is any probability distribution.

3.2 Similarity using Kullback-Leibler divergence

As noted above, we express a probabilistic model and represented a document as a multinomial distribution. To use the probabilistic model in information filtering, we must define a distance between a user profile and a document. Since the distance denote a similarity between the user profile and the document, we propose Kullback- Leibler divergence (KL divegence) to calculate the distance. The KL divergence is a divergence to compare two probability distributions and is defined as below.

$$Div(p_k||q_k) = \sum_{k=1}^{n} q_k \ log\frac{q_k}{q_k} \quad (3)$$

where q_k and p_k are probabilities of kth word in probability distributions of q and p. The KL divergence does not become negative. If two distributions is same, KL divergence is 0.

As the user profile represents user's interest, it is close to the document vector of an interesting document and far from the one of an uninteresting document. Recommending documents according to user's interest, our proposed mnethod selects documents depending on the similarity calculated between a user profile and a document.

3.3 Genetic Algorithm

We make a user profile using a GA since to make the user profile is to find the vector which maximizes the similarity between a gene and an interesting document vector and minimizes the similarity between a gene and an uninteresting document vector. In the GA, a gene represents the user profile and is a real valued vector.

To carry out a GA, a fitness function must be defined. The fitness function is

$$F_k = \sum_{D_i \in I} Div(gene_k||doc_i) - \sum_{D_j \in U} Div(gene_k||doc_j) \quad (4)$$

where $Div(gene_k||doc_i)$ is the KL divergence between a document vector for D_i and the kth gene, I is a set of interesting documents and U is a set of uninteresting documents.

Our crossover is unimodal normal distribution crossover (UNDX) proposed by Ono et al [4]. The UNDX is one of crossover techniques implemented for a real-coded GA and explores a search space depending on a distribution of genes. Pseudocode for the GA is shown in Figuer 1.

4 Experiment

In this section, we describe a test collection, simulation setting and results.

Given initial genes at random. For $t = 0, \cdots, T$:

·Select two genes as parents.

·For $n = 0, \cdots, N$

·Produce two child genes using UNDX.

$\mathbf{C}_1 = \mathbf{m} + z_1\mathbf{e}_1 + \sum_{k=2}^{n} z_k\mathbf{e}_k$

$\mathbf{C}_2 = \mathbf{m} - z_1\mathbf{e}_1 - \sum_{k=2}^{n} z_k\mathbf{e}_k$

$\mathbf{m} = \frac{(\mathbf{P}_1+\mathbf{P}_2)}{2}$

$z_1 \sim N(0, \sigma_1^2)$, $z_k \sim N(0, \sigma_2^2)$

$\sigma_1 = \alpha d_1$, $\sigma_2 = \frac{\beta d_s}{\sqrt{n}}$

$\mathbf{e}_1 = \frac{(\mathbf{P}_1-\mathbf{P}_2)}{|\mathbf{P}_1-\mathbf{P}_2|}$, $\mathbf{e}_i \perp e_j (i \neq j)$

·Select two genes as survivors from all child genes.

Figure 1: The GA algorithm.

4.1 Data

To evaluate a proposed method, we use NTCIR2 test collection which consists of many evaluated documents. In the NTCIR2, many tasks, for example queries on information retrieval, homesickness syndroand and singular point, are prepared. In each task, a document has a label which denotes whether it is relevant or not.

In this experiment, we regard relevant documents as interesting documents and make a user profile. We pick up 14 tasks which have more than 20 relevant documents and use them as experiment tasks. Since a GA is a probabilistic optimizing method, we carry out 10 trials per one task. Traning documents consist of 100 documents drawn at random from all documents and test data consist of the other documents. Each trial uses different training data and test data. In Table 1, we show the number of all documents, all relevant documents and average relevant documents in one trial.

4.2 Simulation and Result

To discuss our proposed mnethod, we compare it with relevance feedback. The relevance feedback is used in the VSM and uses an innar product as a similarity between a user profile and a document. The similarity calculated with an inner product is defined below.

$$sim_i = profile \cdot doc_i \tag{5}$$

where sim_i is a similarity between a user profile and the ith document, doc_i is a document vector of the ith document Doc_i and $\cdot$ denotes an inner product. The document vector is calculated using tf·idf in the relevance feedback. In a comparative experiment, traning data is the same data as in our proposed mnethod.

Table 1: The tasks used in the experiment.

Task No.	all	relevant	relevant/trial
108	371	64	19.0
109	453	21	3.9
110	624	34	5.0
111	540	40	7.2
114	616	20	3.1
115	931	94	9.6
119	387	20	4.6
121	501	63	11.4
126	279	22	6.8
132	210	22	11.2
138	253	24	10.6
139	386	142	39.0
140	309	23	7.7
147	591	80	13.1

Table 2: The Parameters of GA.

population	Generations	Crossover	α	β
5000	50000	20	0.5	0.35

The relevance feedback is defined below.

$$profile = a \sum_{D_i \in I} doc_i - b \sum_{D_j \in U} doc_j \tag{6}$$

where both a and b are arbitrary positive numbers. We decide both a and b using leave-one-out.

Since our proposed mnethod uses a GA, the parameters of the GA for two methods are shown in Table 2. The GA needs more population and generationsto explore wide search space.

In Table 3, 11-point average precision is shown. The 11-point average precision is calculated with 11 precision on 11 recall levels.

4.3 Discussion

In Table 3, we realize that our proposed mnethod is superior to relevance feedback in 9 tasks. However, in the other 5 tasks, relevance feedback is superior to the proposed mnethod. We discuss this 5 tasks in detail especially.

Since a genetic algorithm is probabilistic method, the proposed mnethod depends on various conditions (initial genes, generation of child genes and selection). Hence, we discuss a difference of 11-point average precision using t-test. The t-test is a statistical method to consider whether this difference is significant. In

Table 3: The 11-point average precision.

Task No.	KL divergence($\hat{\mu}_{KL}$)	cosine($\hat{\mu}_{COS}$)
108	0.342	0.282
109	0.189	0.100
110	0.170	0.076
111	0.164	0.188
114	0.152	0.085
115	0.261	0.169
119	0.174	0.087
121	0.390	0.281
126	0.112	0.287
132	0.206	0.208
138	0.609	0.641
139	0.807	0.772
140	0.207	0.192
147	0.357	0.376

Table 4: The result of t-test.

Task No.	t
108	2.937
109	2.329
110	3.599
111	<u>1.198</u>
114	1.942
115	3.809
119	2.667
121	2.765
126	<u>3.547</u>
132	<u>0.064</u>
138	<u>0.664</u>
139	2.020
140	0.430
147	<u>0.530</u>

this case a null hypothesis H_0 is $\mu_{KL} = \mu_{COS}$ where μ_{KL} is an expectation of 11-point average precision for the proposing method and μ_{COS} is an expectation of 11-point average precision for the relevance feedback. First, the difference of the 11-point average precisions is transformed with next formula in each task.

$$t = \frac{\hat{\mu}_{KL} - \hat{\mu}_{COS}}{s\sqrt{\frac{1}{n_{KL}} + \frac{1}{n_{COS}}}} \tag{7}$$

where s is standard deviation of 11-point average precisions for all trials, n_{KL} is the number of 11-point average precision for the proposing method ($n_{KL} = 10$) and n_{COS} is the number of 11-point average precision for the relevance feedback ($n_{COS} = 10$). Since we adopt the level of significance of 5% in this t-test, t becomes 2.101. When t is more than 2.101, the null hypothesis H_0 is rejected. In table 4, we show the result of t-test. When a value of relevance feedback is bigger, the t is underlined.

In Table 4, we find that in 6 tasks the proposed mnethod is superior to the relevance feedback in the view of t-test and in one task inferior.

5 Conclusion

In this paper, we proposed information filtering using a probabilistic model. We introduced a multinomial distribution to express documents and KL divergence to calculate a similarity between a user profile and a document. We confirmed that the proposed mnethod is superior to relevance feedback in some tasks with significance test.

We will apply this proposed mnethod to other situations and evaluate the proposed mnethod in detail.

Acknowledgements

This study is supported by Grants-in-Aid for Scientific Research. We would like to thank MEXT for the supports.

References

[1] G. Salton and M. McGill(1983), Introduction to Modern Information Retrieval, McGrwa-Hill Book Company, New York

[2] C. D. Manning and H. Schutze(1999), Foundation of Statistical Natural Language Processing, MIT Press.

[3] J. H. Holland(1962), Outline for a Logical Theory of Adaptive Systems, Journal of the Association for Computing Machinery, vol.3, pp.297–314.

[4] I. Ono and S. Kobayashi(1997), A Real-coded Genetic Algorithm for Function Optimization Using Unimodal Normal Distribution Crossover, Proceedings of 7th International Conference on Genetic Algorithms, pp.246–253