Information Retrieval Based on Stochastic Models

Masaki Murata Kiyotaka Uchimoto Hiromi Ozaku Hitoshi Isahara

Communications Research Laboratory, Ministry of Posts and Telecommunications

588-2, Iwaoka, Nishi-ku, Kobe, 651-2401, Japan

TEL:+81-78-969-2181 FAX:+81-78-969-2189 http://www-karc.crl.go.jp/ips/murata

{murata,uchimoto,romi,isahara}@crl.go.jp

Abstract

Our method for information retrieval uses Robertson's 2-poisson model which is one type of probablistic approach. This method achieved a comparatively high score in TREC. In the NTCIR contest, we designed systems that could be used for adhoc retrieval and cross-lingual task. This paper explains the systems.

key words 2-poisson model, adhoc retrieval, cross lingual retrieval

確率型手法による情報検索

村田 真樹 内元 清貴 小作 浩美 井佐原 均

郵政省 通信総合研究所 関西先端研究センター 知的機能研究室

〒651-2401 神戸市西区岩岡町岩岡 588-2

TEL:078-969-2181 FAX:078-969-2189 http://www-karc.crl.go.jp/ips/murata

{murata,uchimoto,romi,isahara}@crl.go.jp

概要

われわれの情報検索の方法では基本的に，確率型手法の一つの Robertson の 2- ポアソンモデルを用いている．この方法は，Okapi というシステムで海外で開催された TREC において比較的よい成績を示していたものである．NACSIS のコンテストではこの方法に基づいたシステムをいくつか提出していた．本稿ではこのシステムの説明を行なう．

キーワード 2-poisson モデル，随時検索，言語横断検索

1 Introduction

Information retrieval (IR) has become an increasingly important area of research due to the rapid growth of the Internet. This paper describes our teams' involvement in the NTCIR contest.

In the NTCIR contest[1], our team participated in three tasks: an adhoc retrieval task, a cross lingual retrieval task, and an automatic term recognition task. Here, we describe the adhoc retrieval and cross lingual retrieval tasks. The target database for in-

[1] In the IREX contest, we participated in the named entity and information retrieval tasks [1] [2].

formation retrieval included 330,000 documents extracted from the NACSIS academic conference paper database. It is referred to as the NACSIS test collection. More than half of the documents consisted of pairs of Japanese and English documents. Adhoc retrieval is when the system retrieves the documents from this database that satisfy the condition of a Japanese query. Cross lingual retrieval is when the system retrieves the documents from the database that satisfy the condition of a Japanese query when all the Japanese data have been eliminated and there is only English data.

Our method for information retrieval uses Robertson's 2-poisson model [3] which is one kind of probablistic approach. This method achieved a comparatively high score in TREC. In the following sections, we describe adhoc retrieval and cross lingual retrieval tasks.

2 Adhoc Retrieval

2.1 Outline

The database used in the NTCIR contest includes about 330,000 documents extracted from the NACSIS academic conference paper database. In the adhoc retrieval task, the system retrieves the documents from this database satisfying the condition of a Japanese query.

An example of a Japanese query is shown in Figure 1. (The data is extracted from the preliminary test collection.)

In this figure, the number indicated by "q =" in "retrieval subject" (検索課題) means the number of the queries. In "retrieval request" (検索要求), a phrase that indicates the information needed is written. In "retrieval request explanation" (検索要求説明), the sentences that restrict the information are written. Key words and category information are also given in this figure. In the task, the system receives this data and outputs a document as shown in Figure 2 as a result of the information retrieval.

In the bibliographical data shown in Figure 2, the title of the document is indicated by "TITL," the abstract is indicated by "ABST," and the key words are indicated by "KYWD." In the actual contest we submitted the ID number such as gakkai-0000003395. In the NTCIR, we submit the ID numbers of the 1000 documents for each query. If we submit the results containing correct documents with a higher rank, we can obtain a higher precision. The tool "trec_eval" of TREC is used to evaluate the retrieval result. The tool outputs the two evaluation values of "Average precision" (the average of the precisions for the recall values of 0, 0.1, .., and 1.0) and "R-Precision" (the average of the precisions for the 5, 10, 15, 20, 30, 100, 200, 500, and 1000 documents).

2.2 Outline of Information Retrieval

Our method for information retrieval uses Robertson's 2-poisson model which is one of the probablistic approaches. Robertson's method calculates each document's score in the following equation and outputs the documents having high scores as retrieval

<検索課題 q=0001>

<タイトル>
ロボット
</ タイトル>

<検索要求>
自律移動ロボットについて
</ 検索要求>

<検索要求説明>
自律移動ロボット自体の設計、開発、評価などが総合的に書かれた文献、または、自律移動ロボットにおける部分的なシステム（経路制御、物体認識など）の設計について書かれた文献が検索要求を満たす。自律移動はするがロボットではないものの設計、開発に関する論文も部分的に検索要求を満たす。自律分散システムなどの自律移動ロボットを応用したシステム、自律移動しないロボットに関する文献は検索要求を満たさない。
</ 検索要求説明>

<概念>
自律移動，自律走行，ロボット，画像処理，物体認識，ファジィ制御，カメラ，センサ，ロボットビジョン，ステレオビジョン，経路地図，認知地図，自律分散システム，分散アルゴリズム，協調，設計，開発，評価
</ 概念>

<分野>
1. 電子・情報・制御
</ 分野>

</ 検索課題>

Figure 1: An example of a query

<ACCN>gakkai-0000003395</ACCN>

<TITL TYPE=''kanji''> ライントレース移動ロボットの走行制御: ラインの曲率半径に応じて速度を変える方法 </TITL>

<TITE TYPE=''alpha''>Drive Control of Line Trace Robot</TITE>

<AUPK TYPE=''kanji''> 春日 智恵 / 磯尾 優之 / 野村 民也 / 原島 文雄 </AUPK>

<AUPE TYPE=''alpha''>Kasuga,Chie / Isoo,Masayuki / Nomura,Tamiya / Harashima,Fumio</AUPE>

<CONF TYPE=''kanji''> 平成元年電気学会全国大会一般講演 </CONF>

<CNFD>1989. 04. 04 - 1989. 04. 06</CNFD>

<ABST TYPE=''kanji''><ABST.P> 誘導ラインをトレースする自律型の移動ロボットを高速で走行させるとき、ラインの曲率半径が変化すると、外側車輪の速度入力が一定の場合では、ライン検出と制御信号を出すまでに時間遅れがあるため、曲率半径が小さくなるとラインから脱線してしまう。そこで安定な速度を得るため、曲率半径の異なるラインへの入射角に対するずれ幅を調べ、入射角によらず、また曲率半径が変化しても、ラインからのずれ幅が一定となる速度を決定し、曲率半径に対応した制御信号を出し、より安定した速度で走行す　る方法を提案し、シミュレーションによって安定性の比較、検討を行う。 </ABST.P></ABST>

<KYWD TYPE=''kanji''> 移動ロボット // 走行制御 // 自律型 // CCD センサ </KYWD>

<KYWE TYPE=''alpha''>Robot vehicle // Drive Control // Autonomous // CCD Sensor</KYWE>

<SOCN TYPE=''kanji''> 電気学会 </SOCN>

<SOCE TYPE=''alpha''>The Institute of Electrical Engineers of Japan</SOCE>

Figure 2: An example of the retrieval result

results:

$$Score = \sum_{\text{all terms}} \left(\frac{TF}{k_t \frac{length}{\Delta} + TF} \times log \frac{N}{DF} \right) \quad (1)$$

where terms occur in queries. TF is the frequency of the term in the document, DF is the number of the documents where the term occurs, N is the number of all the documents, $length$ is the number of the characters in the document, Δ is the average of the lengths of all the documents, and k_t is a constant which is set by experiments.

This equation uses a complicated TF numerical term such as $\frac{TF}{\frac{length}{\Delta}+TF}$. The reason is as follows: Even if the value of TF is ∞, the value of the numerical term is at most 1. So all the terms are evaluated uniformly.

2.3 How to extract terms

This section describes how to extract terms. With regard to term extraction we considered several methods, as listed below:

1. A method using only the shortest terms

   This is the simplest method. The method divides the query sentence into short terms by using the morphological analyzer "juman" [4] and eliminates stop words. The remaining words are used in retrieval.

2. A method using all term patterns

   In the first method the terms are too small. For example, "private" and "enterprise" are used instead of "private enterprise." We thought that we should use "private enterprise" in addition to the two shorter terms. Therefore, we decided to use both short and long terms. We call this method "the method using all term patterns." For example, when "Japanese private enterprise" is input, we use "Japanese," "private," "enterprise," "Japanese private," "private enterprise," and "Japanese private enterprise" as terms for information retrieval.

Next, let's see how to use extracted terms in Eq (1).

Query term frequency is used as in the following equation by Robertson:

$$Score = \sum_{\text{all terms}} \left(\frac{TF}{k_t \frac{length}{\Delta} + TF} \times log \frac{N}{DF} \times \frac{TFq}{TFq + kq} \right) \quad (2)$$

where TFq is the frequency of the term in the query, and kq is a constant which is set by experiments.

Similarly, we can also use IDF in the query, and we made the following equation.

$$Score = \sum_{\text{all terms}} \left(\frac{TF}{k_t \frac{length}{\Delta} + TF} \times log \frac{N}{DF} \times \frac{TF_q}{TF_q + k_q} \times log \frac{N_q}{DF_q} \right) \quad (3)$$

where Nq is the number of all the queries and DFq is the number of the queries where the term occurs. The terms which occur in more queries are more

likely to be stop words such as "documents" and "thing." We can decrease the score of stop words by using $log\frac{N_q}{DF_q}$.

2.4 Query expansion by synonyms

The dictionary used in query expansion was made from the database of NACSIS test collection. The procedure has the following steps:

1. We extract pairs of English and Japanese key words from documents in the database.
2. We treat the Japanese key words which have the same English key words of a certain Japanese key word as the synonyms of the Japanese key word.

We expand terms in a query by using this dictionary, and we connect the synonyms of a certain term into one group by using the symbol of "or," and suppose that the a set of synonyms occurs if one of them occurs. Synonyms are less reliable than normal terms, and are weighted by 0.5.

2.5 The method of human term extraction

In this work, we used human term extraction in addition to automatic-term extraction. The procedure of human-term extraction proceeds as follows:

1. We extract all patterns of terms by using "the method using all the term patterns."
2. We make a table that includes each term, its IDF value, and its synonyms.
3. We make a table that shows the results of meaning sort [5] independently of 2. (Meaning sort sorts words by their meaning.) The elimination of stop words is easily done by grouping the terms by meaning sort. (An example of meaning sort is shown in Figure 3.)
4. We make a term list used in information retrieval by using the tables made in 2 and 3. Here, we eliminate terms whose IDF value is high but not effective, and add new terms which are written in the synonym table and are effective.

The information retrieval is done by using Eq (1) with the terms extracted by these methods.

2.6 Experimental Results

We submitted 11 systems in the adhoc retrieval task of the NTCIR contest. The results are shown in Table 1. In the table, "auto" means automatic retrieval, "interact" means retrieval by hands, "short" means short query, and "long" means long query. The columns of length, IDFq, and "Query expansion" indicate these functions are used or not. System D is a tune-up version of System C. All the systems use the method with all the term patterns.

By examining the results, we obtained the following information:

- Because System A is higher than System B, which uses query expansion, query expansion is effective.

[LOCATION] 限界 (limit), 送信者側 (sender)

[QUANTITY] マルチ (multi), 割合 (rate), データ転送レート (data transfer rate), レート (rate), 転送レート (transfer rate), 複数 (multiple)

[TIME etc.] キャスト環境 (cast environment), マルチキャスト環境 (multicast environmen), 環境 (environment), 従来 (past)

[RELATION] 受信者 (receiver), 送信者 (sender), データ (data), 受信データ (received data), 複数データ (multiple data), 相互 (interaction), 存在 (existence), データ品質 (data quality), 品質 (quality), 手動 (manual), 送信者手動 (sender manual)

[Action] 音声 (voice), データ転送レート制御手法 (data transfer rate control method), レート制御手法 (rate control method), 制御手法 (control method), 転送レート制御手法 (transfer rate control method), キャスト通信 (cast communication), マルチキャスト通信 (multicast communication), 通信 (communication), 送信 (send), 受信 (receive), メディア (media), 動画 (motion image), 取り (take), キャスト (cast), データ転送レート制御 (data transfer rate control), データ品質制御 (data quality control), フロー制御 (flow control), レート制御 (rate control), 制御 (control), 転送レート制御 (transfer rate control), 品質制御 (quality control), 扱い (treatment), データ転送 (data transfer), 転送 (transfer)

[UNKNOWN] QOS (QOS), データレート (data rate), フロー (flow), マルチキャスト (multicast), マルチメディア (multimedia) マルチメディアデータ (multimedia data), メディアデータ (media data), 関連性 (relationship), 取り扱い (treatment), 受信データレート (received data rate)

Figure 3: An example of meaning sort

Table 1: Results of adhoc retrieval

	auto or interact	Query Type	TF k_t	length	TFq k_t	IDFq	Query Expansion	Average precision A-judge	Average precision B-judge	R-Precision A-judge	R-Precision B-judge
System A	auto	short	0	—	0	no	no	0.2290	0.2402	0.2419	0.2481
System B	auto	short	0	—	0	no	yes	0.2404	0.2492	0.2490	0.2602
System C	auto	short	1	no	0	no	no	0.2571	0.2646	0.2707	0.2761
System D	auto	short	1	no	0	no	no	0.2575	0.2654	0.2699	0.2753
System E	auto	short	1	yes	0	no	no	0.2584	0.2675	0.2763	0.2910
System F	auto	short	1	yes	0	yes	no	0.2627	0.2732	0.2872	0.3040
System G	auto	long	0	—	0	no	no	0.3337	0.3575	0.3310	0.3665
System H	auto	long	0	—	∞	no	no	0.3716	0.4075	0.3815	0.4122
System I	auto	long	1	no	∞	yes	no	0.3917	0.4323	0.3905	0.4354
System J	interact	long	0	—	0	no	no	0.3479	0.3922	0.3775	0.4160
System K	interact	long	1	no	0	no	no	0.3889	0.4404	0.3956	0.4404

Table 2: Additional experimental results

auto or interact	Query Type	TF k_t	length	TFq k_t	IDFq	Query Expansion	Average precision A-judge	Average precision B-judge	R-Precision A-judge	R-Precision B-judge
auto	long	1	yes	∞	yes	no	0.4182	0.4585	0.4246	0.4672

- When Systems A, D, and E are compared, we can see that Robertson' method is more precise and this is also true for Systems H and I.
- The comparison of Systems C and F indicates that the IDF numerical term is effective.
- $k_q = \infty$ has higher precision than $k_q = 0$ in long query retrieval.
- We extracted terms by hand in System J and K. They did not have the exceptionally higher precision than automatic retrieval systems. Retrieval by hand greatly increases the human cost and was found to be ineffective.

2.7 Additional experiments

We carried out additional experiments after the NTCIR contest. The result is shown in Table 2. After the contest, we noticed that the method using all term patterns was not effective and the method using only the shortest terms was effective. So the experiments in Table 2 use the method which uses only the shortest terms. The result is about 0.02 higher than System I.

<ACCN>gakkai-0000003395</ACCN>
<TITE TYPE=''alpha''>Drive Control of Line Trace Robot</TITE>
<AUPE TYPE=''alpha''>Kasuga,Chie / Isoo,Masayuki / Nomura,Tamiya / Harashima,Fumio</AUPE>
<CNFD>1989. 04. 04 - 1989. 04. 06</CNFD>
<KYWE TYPE=''alpha''>Robot vehicle // Drive Control // Autonomous // CCD Sensor</KYWE>
<SOCE TYPE=''alpha''>The Institute of Electrical Engineers of Japan</SOCE>

<ACCN>gakkai-0000010478</ACCN>
<TITE TYPE="alpha">An Visual image processing system for an autonomous vehicle</TITE>
<AUPE TYPE="alpha">Ozaki,Tohru /Ohzora,Mayumi / Hiratsuka,Yoshitaka</AUPE>
<CNFE TYPE="alpha">The Special Interest Group Notes of IPSJ</CNFE>
<CNFD>1990. 11. 22</CNFD>
<ABSE TYPE="alpha"><ABSE.P>We have
developed a high speed image processing system for an autonomous
vehicle PVS,Personal Vehicle System.The goal of PVS was to be able to
run autonomously,detecting white lines and obstacles on a road
including straightsections,curves,and
intersections.</ABSE.P><ABSE.P>A video-rate image processing system
was developed and utilized in order to realize high-speed image
processing.Position data for white lines and for obstacles is
obtained at every 30 milliseconds.</ABSE.P><ABSE.P>Exactly,the PVS
was able to run autonomously at 60 km, h on straight sections,at 15
km/h on curves,and at 5 km/h through intersections in our testing
ground.</ABSE.P></ABSE>
<KYWETYPE="alpha">autonomous vehicle // image processing // white line detection //
obstacle detection // stereo images</KYWE>
<SOCE TYPE="alpha">Information Processing Society of Japan</SOCE>

Figure 4: Examples of the retrieval result

3 Cross lingual retrieval

3.1 Outline

The database used in cross lingual retrieval is the database which eliminates the Japanese data and only includes English data. English information can be retrieved by a Japanese query. Examples of retrieval results are shown in Figure 4. Queries are equal to those in adhoc retrieval.

3.2 How to retrieve

In cross lingual retrieval, queries and documents can be translated by machine translation systems. In our system, we translated the queries. Our cross lingual retrieval is performed by using the same method as adhoc retrieval without using translation.

In our system we translate Japanese terms into English terms by using a word translation dictionary. We use the following two methods for Japanese term extraction.

1. juman only

   We divide a query into words by the morphological analyzer "juman" and eliminate stop words.

2. all the patterns

   We obtain Japanese terms by using "the method using all term patterns" described in Section 2.3.

We used the following two dictionaries for translation into English.

1. EDR Japanese-English Bilingual Dictionary [6].

2. A Japanese-English bilingual dictionary made from pairs of English and Japanese key words in the database of the NACSIS test collection.

   This dictionary is made by treating the English key words of a certain Japanese key word as the English translation of the Japanese key word. However, because this method uses the Japanese data from the NACSIS test collection, it may not satisfy the condition of cross lingual retrieval.

When a Japanese word has several equivalents in English, we use all the equivalents as in query expansion in Section 2.4.

3.3 Experimental results

We submitted three systems in cross lingual task of NTCIR. The result is shown in Table 3.

In this table, "extract term method" indicates whether the system uses the method of "juman only" or the method of "all the patterns." And "dictionary" indicates whether the system uses the EDR dictionary or the dictionary made from the NACSIS database. In the case of "juman only", we set the parameters in eq (1) and (2) as $k_t = 0$, $k_q = 0$, and do not use IDFq and query expansion. In the case of "all the patterns", we set the parameters in eqs (1) and (2) as $k_t = 1$, $k_q = 0$, and do not use IDFq and query expansion.

We can obtain the following information from the

Table 3: Result in cross lingual retrieval task

	auto or interact	query type	extract term method	dictionary	Average precision		R-Precision	
					A-judge	B-judge	A-judge	B-judge
System L	auto	short	juman only	NACSIS	0.0940	0.0965	0.0908	0.1043
System M	auto	short	juman only	EDR	0.0403	0.0533	0.0583	0.0704
System N	auto	short	all the patterns	EDR	0.0507	0.0543	0.0559	0.0695

table.

1. The dictionary made from the database of NACSIS obtained a higher precision rate than the EDR dictionary. However, this method has problems using Japanese data of the NACSIS test collection.

2. We conducted two experiments with the EDR dictionary: "juman only" and "all the patterns." And "all the patterns" was more precise than "juman only."

4 Conclusion

This paper described our systems based on Robertson's 2-poisson model, which is one type of probablistic approach. We believe that the test collection made in this contest is very valuable. In the future we would like to study the systems designed by the other teams to further our research in the area of information retrieval.

Acknowledgments

We would like to thank the staff and participants at the NTCIR contest. This work would not have been possible without their help.

Reference

[1] Kiyotaka Uchimoto, Masaki Murata, Hiromi Ozaku, and Qing Ma. Named entity extraction based on maximum entropy model and transformation rules — evaluation in irex-ne formal run —. *Proceedings of the IREX Workshop*, 1999.

[2] Masaki Murata, Kiyotaka Uchimoto, Hiromi Ozaku, and Qing Ma. Information retrieval based on stochastic models in irex. *Proceedings of the IREX Workshop*, 1999.

[3] S. E. Robertson S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retreival. In *In Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994.

[4] Sadao Kurohashi and Makoto Nagao. *Japanese Morphological Analysis System JUMAN version 3.5.* Department of Informatics, Kyoto University, 1998. (in Japanese).

[5] Masaki Murata, Kyoko Kanzaki, Kiyotaka Uchimoto, Qing Ma, and Hitoshi Isahara. Meaning sort msort — three examples: Dictionary construction, tagged-corpus construction, and information presentation system —. *Information Processing Society of Japan, WGNL 130-12*, 1999.

[6] EDR (Japan Electronic Dictionary Research Institute, Ltd.). *EDR Electronic Dictionary Technical Guide*, 1993.