Cyclone: An Encyclopedic Web Search Site

Atsushi Fujii
Graduate School of Library, Information and Media Studies
University of Tsukuba
1-2 Kasuga, Tsukuba, 305-8550, Japan
fujii@slis.tsukuba.ac.jp

Katunobu Itou
Graduate School of Information Science
Nagoya University
Furo-cho, Chikusa-ku, Nagoya-shi, 464-8603, Japan
itou@is.nagoya-u.ac.jp

Tetsuya Ishikawa
Graduate School of Library, Information and Media Studies
University of Tsukuba
1-2 Kasuga, Tsukuba, 305-8550, Japan
ishikawa@slis.tsukuba.ac.jp

ABSTRACT

We propose a Web search site called "CYCLONE", in which a user can retrieve encyclopedic term descriptions on the Web. CYCLONE searches the Web for headwords and page fragments describing the headwords. High-quality page fragments are selected as term descriptions and are classified into domains. The number of current headwords is over 700,000.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.5 [**Information Storage and Retrieval**]: Online Information Services; I.2.7 [**Artificial Intelligence**]: Natural Language Processing

General Terms

Experimentation, Management

Keywords

Encyclopedias, Web Search, Extraction, Organization

1. INTRODUCTION

The World Wide Web, which contains an enormous volume of up-to-date information, is a promising source to obtain encyclopedic knowledge. It has become common to consult the Web for specific keywords, instead of consulting conventional encyclopedias. However, existing Web search engines often retrieve extraneous pages containing low-quality, unreliable, and misleading information.

Fujii and Ishikawa [1] proposed an automatic method to extract term descriptions from the Web and classify multiple descriptions into domains and word senses. Using this method, we propose a Web search site called "CYCLONE", in which a user can efficiently obtain an encyclopedic term description in a specific word sense.

2. DESCRIPTION OF CYCLONE

Figure 1 depicts the overall design of CYCLONE, which produces a corpus off-line. Users search the resultant corpus for specific term descriptions on-line. More than 700,000 Japanese terms are currently indexed as headwords.

WWW 2005, May 10–14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

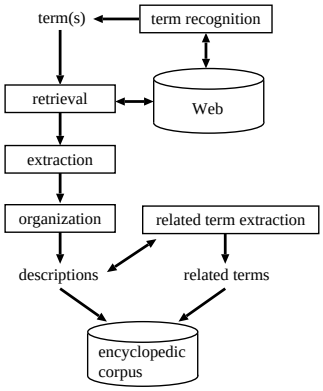

Figure 1: Overall design of CYCLONE.

In the off-line process, the term recognition module periodically searches the Web for new morpheme sequences, which are used as target terms. The retrieval module searches the Web for pages including a target term. The extraction module analyzes the layout (i.e., the structure of HTML tags) of the retrieved pages and identifies paragraphs that potentially describe the target term. While promising descriptions are extracted from pages resembling on-line dictionaries, descriptions can also be extracted from other types of pages, such as blogs.

The organization module classifies multiple paragraphs for the term into predefined domains (e.g., computers and medicine) and sorts them according to a probability score. Different word senses, which are often associated with different domains, are distinguished and high-quality descriptions are selected for each domain. The probability that paragraph d is selected as a description for domain c, $P(d|c)$, is transformed as in Equation (1), by the Bayesian theorem.

$$P(d|c) = \frac{P(c|d) \cdot P(d)}{P(c)} \quad (1)$$

$P(c|d)$ models the probability that d corresponds to c. $P(d)$ models the probability that d is a description for the target term, disregarding the domain. We shall call them domain and description models, respectively. We regard $P(c)$ as a constant. $P(c|d)$ is modeled for 22 domains by a statistical categorization method. We decompose $P(d)$ into language, reliability, and layout properties, as shown in Equation (2).

$$P(d) = P_L(d) \cdot P_R(d) \cdot P_S(d) \quad (2)$$

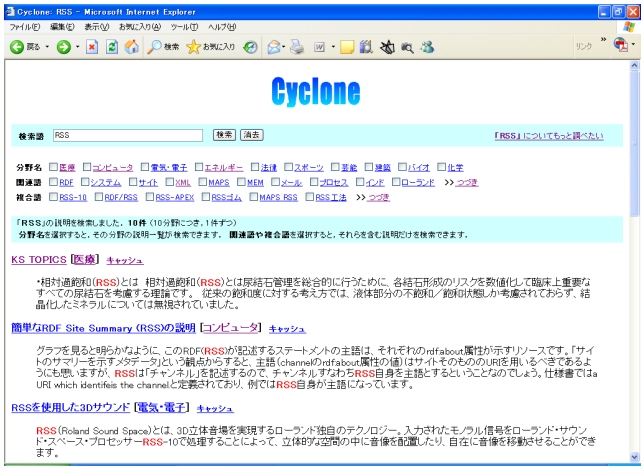

Figure 2: Example descriptions for "RSS".

$P_L(d)$, $P_R(d)$, and $P_S(d)$ denote language, reliability and layout (structure) models, respectively. $P_L(d)$ is a trigram language model produced from a machine readable encyclopedia. Unlike our previous method [1], $P_R(d)$ and $P_S(d)$ are proposed in this paper. For P_R, we implemented a software to compute PageRank, which is used in Google[1] to rate the quality of Web pages based on hyperlink information. $P_R(d)$ is the PageRank value for the page from which d was extracted. If d is extracted from a page whose HTML layout is similar to one typically used to describe terms, $P_S(d) = 1$. Otherwise, $P_S(d) = 0.5$. The HTML layout for a page is obtained in the extraction module.

Finally, the related term extraction module searches top-ranked descriptions for terms related to the target term.

In the on-line process, users can access the corpus by a number of methods. A simple method is to retrieve one or more descriptions for a submitted term. Figure 2 depicts an example retrieval result in response to the term "RSS"[2]. In the half bottom of Figure 2, three descriptions classified into different domains, that is, "Relative SuperSaturation" (medicine), "RDF site summary" (computers), and "Roland Sound Space" (electricity), are retrieved. Below the input box, automatically extracted related terms (e.g., "RDF" and "XML") are displayed, which can be used as feedback terms to narrow down the user focus.

If a submitted term is not indexed as a headword, headwords which share substrings with the submitted term are proposed. This method is effective for variants and misspelling. Headwords which share an English translation with the submitted term are also proposed. These headwords are often synonyms of the submitted term. The corpus can also be queried by WH-questions, such as "who invented the printing press?" and "what is the capital of Canada?" In addition, multiple descriptions for a term can be summarized to produce a condensed single description [2].

3. EVALUATION

In this paper, we focus only on evaluating the organization module. We collected test terms from the index of a printed terminology dictionary, which lists 2226 technical terms frequently appearing in the Information Technology Engineers Examinations. We performed two experiments using 2080 terms for which at least one paragraph was obtained.

[1]http://www.google.com/
[2]http://cyclone.slis.tsukuba.ac.jp/

Table 1: Effectiveness of sorting paragraphs.

	R	RS	RL	RSL
MAP	.204	.247	.410	.433
MRR	.280	.436	.595	.639
RANK	28.6	21.3	9.7	7.5

In the first experiment, we evaluated the effectiveness of the reliability, layout, and language models in sorting paragraphs. For each test term, paragraphs were sorted according to PageRank and at most top 500 paragraphs were manually judged as to whether or not it is a correct description for a term in question. The average number of paragraphs judged per term was 141. In addition, each correct paragraph was manually annotated with one or more domains.

We used three evaluation measures. First, we used mean average precision (MAP), which is a combination of recall and precision and has commonly been used to evaluate information retrieval. MAP becomes great if many correct paragraphs are sorted in high ranks for each test term. This measure is important, if a user requires more than one correct description for a single term.

Second, we used mean reciprocal rank (MRR), which has commonly been used to evaluate question answering. For each test term, we calculate the reciprocal of the rank at which the first correct paragraph was found. MRR is the mean of the reciprocal ranks for all test terms. This measure is important, if a user requires only one correct description. Third, we used the average rank at which the first correct paragraph was found. Unlike MRR, this value, which we shall call "RANK", is in proportion to the rank. While greater MAP and MRR are obtained by a better method, smaller RANK is obtained by a better method.

Table 1 shows MAP, MRR, and RANK for different combinations of the reliability (R), layout (S), language (L) models. "R" simulates a conventional search engine and is a baseline. The layout and language models were independently effective and when used together the improvement was even greater, disregarding the evaluation measure.

In the second experiment, we used 1472 of the test terms to which one or more correct descriptions and domains were manually annotated and evaluated the effectiveness of the domain model in categorizing paragraphs for these terms. We regarded only the top domain determined by the domain model as the system output. The recall and precision were 0.671 and 0.700, respectively. Thus, approximately 70% of correct descriptions can be found in correct domains.

4. CONCLUSION

We proposed the CYCLONE search site, which automatically produces an encyclopedic corpus on the Web and provides users with a number of access methods. Future work includes evaluating CYCLONE on real users' behaviors.

5. REFERENCES

[1] A. Fujii and T. Ishikawa. Organizing encyclopedic knowledge based on the Web and its application to question answering. In *Proc. of ACL 2001*, pages 196–203, 2001.

[2] A. Fujii and T. Ishikawa. Summarizing encyclopedic term descriptions on the Web. In *Proc. of COLING 2004*, pages 645–651, 2004.