SNUMedinfo at ImageCLEF 2013: Medical retrieval task

Sungbin Choi, Jeongeun Lee, Jinwook Choi

Medical Informatics Laboratory, Seoul National University, Seoul, Republic of Korea

wakeup06@empas.com, jeleedict@gmail.com, jinchoi@snu.ac.kr

Abstract. This paper describes the participation of the SNUMedinfo team at the two retrieval tasks (Ad-hoc image-based retrieval and Case-based retrieval) in the ImageCLEF 2013 medical task.
For the ad-hoc image-based retrieval task, we submitted 1 baseline textual run using query likelihood model in Indri search engine, and 4 visual runs utilizing various image features implemented in Lire image retrieval library.
For the case-based retrieval task, we submitted 1 baseline textual run using query likelihood model in Indri search engine, and 9 textual runs utilizing external corpus (MEDLINE) for expansion term inference. Our method ranked first in the case-based retrieval task.

Keywords: Case-based retrieval, Query expansion, Inference, MEDLINE, MeSH, Language model, Medical information retrieval, Indri, Lire

1. Introduction

In this paper, we describe the methods in participation of the two retrieval tasks (Ad-hoc image-based retrieval and Case-based retrieval task) in the ImageCLEF 2013 medical task. For detailed explanation about each task's characteristics, please see [1].

2. Ad-hoc image-based retrieval

2.1 Textual run

We submitted 1 baseline run (SNUMedinfo11) using unigram language model with Dirichlet prior smoothing[2, 3]. Only figure caption field in document were indexed. Experimental results are described in Table 1.

Table 1. Textual run result in ad-hoc image-based retrieval task

Runid	MAP	GM-MAP	bpref	P10	P30
SNUMedinfo11	0.18	0.0266	0.1866	0.2657	0.1895

2.2 Visual run

We submitted 4 visual runs using various features implemented in Lire image retrieval library[4], respectively. These features are : 1) the auto color correlogram feature which, as the faster and more efficient version of the color correlogram[5] uses the color of immediate neighborhood, 2) the color and edge directivity descriptor(CEDD) feature[6] which creates a 24-bin fuzzy color histogram in HSV color space, 3) the fuzzy color and texture histogram(FCTH) feature[7] which uses the same color scheme with CEDD but describes edge information extensively, 4) the joint composite descriptor(JCD)[8] which is the combination of CEDD and FCTH.

There are several query images given per each query. We draw ranked list per each query image, and then combined them using Borda-fuse methods[9].

Experimental results are described in Table 2.

Table 2. Visual run result in ad-hoc image-based retrieval task

Runid	MAP	GM-MAP	bpref	P10	P30
SNUMedinfo12	0.0033	0.0001	0.0153	0.0257	0.0219
SNUMedinfo13	0.0043	0.0002	0.0126	0.0229	0181
SNUMedinfo14	0.0023	0.0002	0.009	0.0171	0.0124
SNUMedinfo15	0.0019	0.0002	0.0074	0.0086	0.0114

3. Case-based retrieval

3.1 Textual run

We submitted 1 baseline run (SNUMedinfo10) using unigram language model with Dirichlet prior smoothing. Title, abstract and fulltext field in document are indexed. The queries are stopped at the query time using the standard 418 INQUERY stopword list, case-folded, and stemmed using Porter stemmer.

For other 9 textual runs (SNUMedinfo1~9), we utilized external corpus (MEDLINE) for robust and effective expansion term inference. We leased 2013 MEDLINE/PubMed Journal Citations from the U.S. National Library of Medicine, composed of roughly 22 million MEDLINE citations. Our method can be summarized in the following steps (*k* and *w* is parameter).

(1) External corpus (MEDLINE) documents are indexed. Title, abstract and Mesh descriptors fields are indexed.

(2) Using original query, MEDLINE documents were retrieved.
    a. Among retrieved MEDLINE documents, *k* documents whose publication type is "*Case Reports*", are selected from the top rank.
    b. In selected *k* documents, MeSH descriptors whose "*MajorTopicYN*" attribute value is "*Y*" are collected for expansion term.
(3) Original query is augmented with expansion terms. Expansion terms are weighted by parameter *w*.
(4) Expanded query is applied on target corpus.

Indri query example is described below.

#weight (
 (*1-w*) #combine(*original query terms*)
  w #combine(*expansion query terms*))

#weight (
0.7 #combine(48 old woman right cheek swelling blocked nasal passage …)
0.3 #combine(maxillary neoplasms myxoma odontogenic tumors …))

We tried diverse parameter combinations (*k* : 1, 5, 10; *w* : 0.15, 0.3, 0.45) to evaluate the effect of parameter value on the performance. Experimental results are described in Table 3.

Table 3. Textual run result in case-based retrieval task

Runid	MAP	GM-MAP	bpref	P10	P30
SNUMedinfo1	0.221	0.1208	0.1952	0.2343	0.1619
SNUMedinfo2	0.2197	0.0996	0.1861	0.2257	0.1486
SNUMedinfo3	0.1751	0.0606	0.1572	0.2114	0.1286
SNUMedinfo4	0.2228	0.1281	0.2175	0.2343	0.1743
SNUMedinfo5	0.2388	0.1266	0.2259	0.2543	0.1857
SNUMedinfo6	0.2374	0.1112	0.2304	0.2486	0.1933
SNUMedinfo7	0.2172	0.1266	0.2116	0.2486	0.1771
SNUMedinfo8	0.2389	0.1279	0.2323	0.2686	0.1933
SNUMedinfo9	0.2429	0.1163	0.2417	0.2657	0.1981
SNUMedinfo10	0.1827	0.1146	0.1749	0.2143	0.1581

Overall, our methods brought 20~30% performance gain (in MAP) compared to the baseline method, and ranked first in the case-based retrieval task.

4. Conclusion

For the ad-hoc image-based retrieval task, we submitted 4 visual runs using several basic image features without applying additional techniques. These runs showed unsa-

tisfactory performance (MAP range from 0.0019 to 0.0043), which means more advanced technique is required for this task.

For the case-based retrieval task, we submitted 9 textual runs (SNUMedinfo1~9) utilizing external corpus (MEDLINE) for expansion term inference. This method showed effective performance (20~30% gain in MAP), and is also robust across diverse parameter values.

5. Acknowledgements

This work was supported by the Seoul National University Brain Fusion Program Research Grant.

6. References

1. A. Garcia Seco de Herrera, J.K.-C., D. Demner Fushman, S.Antani, H. Muller, *Overview of the ImageCLEF 2013 medical tasks*, in *Working notes of CLEF 2013*2013: Valencia, Spain.
2. Strohman, T., et al. *Indri: A language model-based search engine for complex queries*. in *Proceedings of the International Conference on Intelligent Analysis*. 2005.
3. Zhai, C. and J. Lafferty, *A study of smoothing methods for language models applied to Ad Hoc information retrieval*, in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*2001, ACM: New Orleans, Louisiana, USA. p. 334-342.
4. Lux, M. and O. Marques, *Visual Information Retrieval using Java and LIRE.* Synthesis Lectures on Information Concepts, Retrieval, and Services, 2013. **5**(1): p. 1-112.
5. Jing, H., et al. *Image indexing using color correlograms*. in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*. 1997.
6. Chatzichristofis, S. and Y. Boutalis, *CEDD: Color and Edge Directivity Descriptor: A Compact Descriptor for Image Indexing and Retrieval*, in *Computer Vision Systems*, A. Gasteratos, M. Vincze, and J. Tsotsos, Editors. 2008, Springer Berlin Heidelberg. p. 312-322.
7. Chatzichristofis, S.A. and Y.S. Boutalis. *FCTH: Fuzzy Color and Texture Histogram - A Low Level Feature for Accurate Image Retrieval*. in *Image Analysis for Multimedia Interactive Services, 2008. WIAMIS '08. Ninth International Workshop on*. 2008.
8. Chatzichristofis, S.A., Y.S. Boutalis, and M. Lux. *Selection of the proper compact composite descriptor for improving content based image retrieval*. in *Proceedings of the 6th IASTED International Conference*. 2009.

9. Aslam, J.A. and M. Montague, *Models for metasearch*, in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*2001, ACM: New Orleans, Louisiana, USA. p. 276-284.