Investigation on Application of Local Cluster Analysis and Part of Speech Tagging on Persian Text

Amir Hossein Jadidinejad
Computer Engineering Department,
Islamic Azad University,
Qazvin, Iran.
amir@jadidi.info

Mitra Mohtarami
Database Research Group,
University of Tehran,
Tehran, Iran
m.mohtarami@yahoo.com

Hadi Amiri
Database Research Group,
University of Tehran,
Tehran, Iran
h.amiri@ece.ut.ac.ir

Abstract

In this research we applied Local Cluster Analysis (LCA) in tandem with Part-of-Speech tagging to monolingual task. We study different Persian POS tags and select a set of designated tags to reduce the size of our index and store the rich content of the documents. In addition, we applied LCA on the retrieved documents to detect the relevant and irrelevant documents to the user query. The clustering method is an important part in our approach. So we address the problem of building effective and meaningful clustering and evaluate different well-known and state of the art clustering methods for better efficiency and effectiveness in the proposed approach.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [**Database Management**]: Languages-*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Persian Text Retrieval, Search Result Clustering, Query-Specific Clustering, Local Cluster Analysis, Persian Part-of-Speech Tagging.

1 Pre-processing: Part-of-Speech Tagging

Part-of-speech tagging is the task of annotating each word in a text with its most appropriate syntactic category. Having an accurate POS tagger is useful in many information related applications such as information retrieval, information extraction, text to speech systems, linguistic analysis, etc. To study the effect of POS tags on Persian text retrieval we used a set of Persian POS tags from [21], [22] and based on our observation we selected a set of tags as most meaningful tags: Single and Plural Nouns (N), Adjectives (ADJ), Verbs (V) and Adverbs (ADV). We tagged the Hamshahri collection using TNT tagger [] and for each document in the collection we just keep the terms that have one of the above tags and remove the other terms. Doing so, we try to reduce the index size while we keep the important content of the documents.

Table 1 depicts the precision-recall on the training set when we index documents with different tag sets. As it is shown in Table 2 the mean average precision is definitely improves when we consider the four tags (ADJ/N/V/ADV). In addition, column three shows that the adverbs have not big impact on the retrieval precision and column two confirm that nouns and adjectives have big impact on the retrieval precision.

Table 1 Part-of-Speech tagging results on train set.

Recall	Precision			
	N	ADJ/N	ADJ/N/V	ADJ/N/V/ADV
0.0	0.6915	0.7687	0.7745	***0.7905***
0.1	0.5666	0.6478	0.6547	***0.6573***
0.2	0.5006	0.5789	0.5905	***0.5908***
0.3	0.4234	0.5153	0.5207	***0.5230***
0.4	0.3669	0.4572	0.4718	***0.4716***
0.5	0.3269	0.4093	0.4272	***0.4276***
0.6	0.2691	0.3365	0.3447	***0.3452***
0.7	0.1861	0.2458	0.2623	***0.2629***
0.8	0.1445	0.1928	0.1960	***0.2044***
0.9	0.0624	0.0900	0.0918	***0.0917***
1.0	0.0381	0.0492	0.0498	***0.0492***
MAP	0.3011	0.3685	0.3770	***0.3791***
GMAP	0.1829	0.2892	0.3028	***0.3053***
R-PREC	0.3579	0.4205	0.4202	***0.4212***

2 Post-processing: Local Cluster Analysis

The LCA framework operates as follows (Fig. 1): First initial results retrieve per query based on standard method, then clustering is apply on initial results and separate it into two clusters. After the clustering step, we have to choose relevant cluster and then re-rank results based on it. The proposed architecture has some key features [1]:

- *Simple and high performance*. [1] shows that it's better than the best known standard Persian retrieval systems [3], [5], [14].
- *Independent of initial system architecture*. It can embed in any fabric information retrieval system. It cause proposed architecture very good envisage for the web search engines.
- *High-Precision*. Relevant documents exhibit at top of the result list.

[1] indicate that the LCA technique is effective and efficient with an overall performance superior than best methods in initial retrieval [3], [5] and [14]. There are related works such as [17], [18] and [6]. Following contains brief description about LCA approach, see [1] for more details.

a) The initial retrieval

Some experimental results [2], [3], [5], [12] and [14] show that 4-gram and term based vector space model with Lnu.ltu weighting scheme has acceptable performance for Persian text retrieval so far. The effectiveness of these methods describe in the initial column of Table 2 and Table 3. We leverage Lemur toolkit [12] in this section.

b) Construction of clusters

We consider algorithms that assume the vector space representation for documents and modeled as feature-object matrices (especially term-document matrix).

K-means [4], [20] is probably the most celebrated and widely used clustering technique; hence it is the best representative of the class of iterative centroid-based divisive algorithms. On the other hand, PDDP [7] is representative of the non-iterative techniques based upon the Singular Value Decomposition (SVD) of a matrix built from the data set. PDDP can be quite efficient in comparison to other agglomerative hierarchical algorithms [8]. The authors in [20] presented a comparative analysis on the bisecting k-means and PDDP clustering algorithms.

Two well-known disadvantages of the k-means algorithm are that the generated clusters depend on the specific selection of initial centroid and that the algorithm can be trapped at local minima of the objective function [10]. Therefore, one run of k-means can easily lead to clusters that are not satisfactory and users are forced to initialize and run the algorithm multiple times.

Regarding the PDDP, despite the convenient deterministic nature, it is easy to construct examples where PDDP produces inferior partitioning than k-means [10]. PDDP is known as effective clustering method for text mining [8] [9] when term-document matrix is very large and extremely sparse.

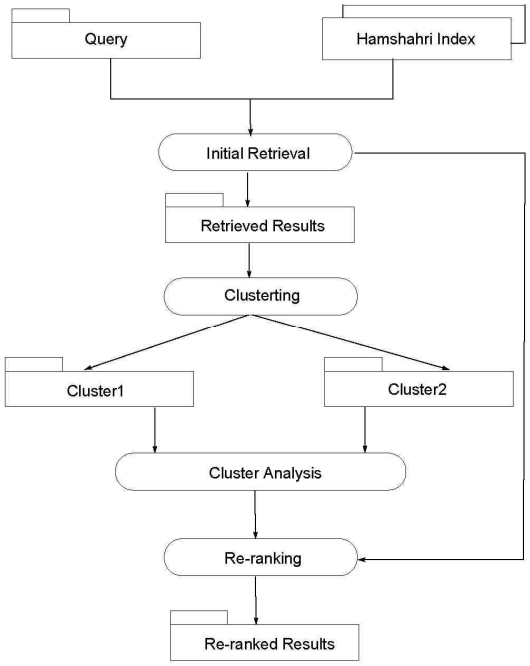

Fig. 1 Local Cluster Analysis architecture

In LCA approach we need deterministic clustering algorithm with high quality semantic, so we turn to some state of the art researches [10] that have been studying the characteristics of PDDP and have been considering ways to improve its performance. [10] shows how to leverage the power of k-means and some interesting recent theory in order to better steer the partitioning decision at each iteration of PDDP.

We apply above algorithms and evaluate the final results per clustering method. TMG [11] has been used for the construction of term-document matrix and used logarithmic local term and IDF global weighting on Hamshahri queries. See results in Table 2 and Table 3.

c) Cluster analysis

In the cluster analysis step we have to analyze clusters content and choose relevant and irrelevant cluster. It's an important selection.

Each cluster has a cluster centroid in the form of a vector which is useful as a representative of a cluster. We conjecture that relevant cluster centroid must be near than irrelevant cluster centroid to the query so clusters centroid and query vector compare with cosine similarity measure [4] and choose relevant cluster.

d) Documents re-ranking

We focus on initial retrieved documents and combine it with clusters evidence. Re-ranked list consist of two sections. Relevant section contains documents in the relevant cluster and the irrelevant section contains documents in the irrelevant cluster in order of initial retrieved documents.

Table 2 Interpolated Recall-Precision on Hamshahri with the best initial retrieval method and different clustering variants. * Average over 100 runs.

Recall	Precision						
	Initial	**PDDP**	**E.K-means***	**S.K-means***	**PDDP_2Means***	**PDDP_OPT_2MEANS**	**PDDP_OPTCUT_PD**
0.0	0.5037	0.8345	0.833500	0.820945	0.845005	0.8284	0.8413
0.1	0.4103	0.6603	0.725994	0.684740	0.718321	0.6992	0.6654
0.2	0.3817	0.5663	0.661786	0.607776	0.657949	0.5774	0.5776
0.3	0.3653	0.5230	0.600378	0.547230	0.603402	0.5096	0.5060
0.4	0.3489	0.4681	0.539994	0.493411	0.549842	0.4586	0.4687
0.5	0.3314	0.4154	0.478150	0.445955	0.497103	0.4185	0.4209
0.6	0.3139	0.3703	0.420919	0.396584	0.445274	0.3732	0.3686
0.7	0.2886	0.3241	0.365966	0.348759	0.389229	0.3330	0.3250
0.8	0.2533	0.2801	0.303962	0.289828	0.321746	0.2806	0.2746
0.9	0.1612	0.1799	0.189509	0.183478	0.200872	0.1740	0.1735
1.0	0.0342	0.0417	0.043988	0.040782	0.045484	0.0362	0.0363

Table 3 Other evaluation measures on Hamshahri with the best initial retrieval method and different clustering variants. * Average over 100 runs.

Criterion	Values						
	Initial	**PDDP**	**E.K-means***	**S.K-means***	**PDDP_2Means***	**PDDP_OPT_2MEANS**	**PDDP_OPTCUT_PD**
MAP	0.2766	0.3957	**0.451731**	0.416990	**0.459822**	0.3982	0.3949
GMAP	0.2186	0.3060	**0.360739**	0.332635	**0.369182**	0.3155	0.3112
R-Prec	0.2898	0.3822	**0.454330**	0.414220	**0.465161**	0.3903	0.3866
P5	0.2646	0.6000	0.636797	0.583229	0.618098	0.5908	0.6000
P10	0.2800	0.5185	0.577830	0.522268	0.560636	0.5077	0.5077
P15	0.2974	0.4800	0.534447	0.487479	0.523282	0.4749	0.4708
P20	0.3023	0.4469	0.505373	0.459854	0.503223	0.4385	0.4369

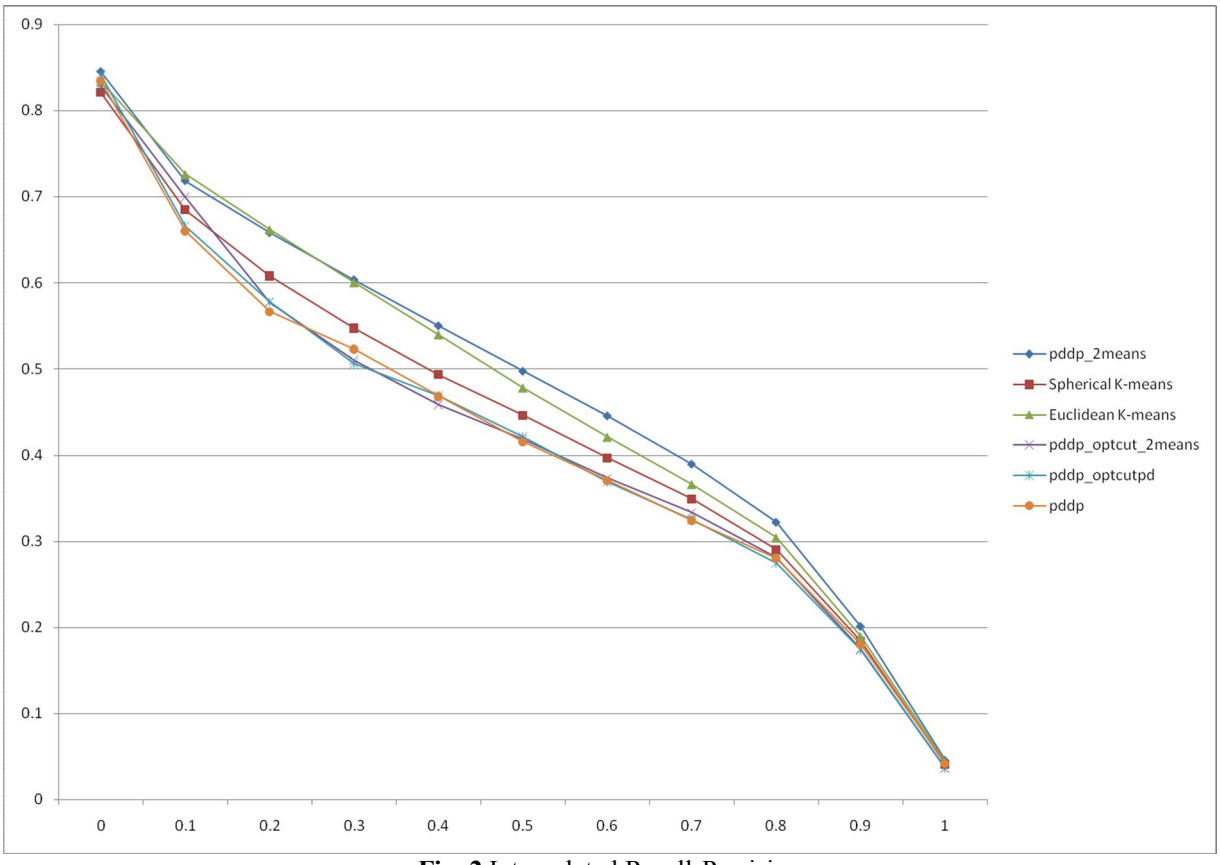

Fig. 2 Interpolated Recall-Precision

Table 4 Different runs on Test Set

Recall	Precision												
	PDDP_2M1	**K-M1**	**PDDP_2M2**	**K-M2**	**PDDP_2M3**	**K-M3**	**PDDP_2M4**	**K-M4**	**PDDP_2M5**	**K-M5**	**PDDP_2M6**	**PDDP**	**INIT**
0.0	0.8146	0.8255	0.8297	0.8204	0.8125	0.8127	0.8181	0.8234	0.8089	0.8341	0.8187	0.8340	0.8320
0.1	0.6351	0.6443	0.6504	0.6457	0.6334	0.6358	0.6383	0.6460	0.6334	0.6512	0.6438	0.6519	0.6390
0.2	0.5526	0.5568	0.5612	0.5621	0.5521	0.5618	0.5509	0.5650	0.5560	0.5719	0.5495	0.5486	0.5595
0.3	0.4817	0.4854	0.4911	0.4907	0.4838	0.4862	0.4797	0.4943	0.4821	0.4957	0.4800	0.4793	0.4870
0.4	0.2985	0.3067	0.3041	0.3055	0.3007	0.3030	0.2986	0.3042	0.3009	0.3147	0.2963	0.2869	0.3065
0.5	0.1853	0.1910	0.1880	0.1925	0.1897	0.1877	0.1870	0.1879	0.1875	0.1947	0.1824	0.1669	0.1864
0.6	0.0981	0.1051	0.1024	0.1112	0.1025	0.1036	0.1039	0.1063	0.1010	0.1084	0.0999	0.0768	0.1030
0.7	0.0789	0.0864	0.0813	0.0900	0.0809	0.0842	0.0815	0.0761	0.0818	0.0789	0.0708	0.0576	0.0838
0.8	0.0447	0.0439	0.0497	0.0487	0.0488	0.0425	0.0492	0.0412	0.0478	0.0408	0.0472	0.0367	0.0510
0.9	0.0297	0.0282	0.0325	0.0307	0.0317	0.0283	0.0319	0.0264	0.0318	0.0285	0.0324	0.0295	0.0354
1.0	0.0210	0.0180	0.0212	0.0216	0.0174	0.0180	0.0174	0.0199	0.0171	0.0178	0.0212	0.0173	0.0241
MAP	0.2637	0.2705	0.2711	0.2746	0.2664	0.2687	0.2661	0.2708	0.2659	**0.2761**	0.2642	0.2591	0.2718
GMAP	0.1994	0.2061	0.2068	0.2018	0.2030	0.2050	0.2040	0.2064	0.2028	**0.2080**	0.2013	0.2026	0.2056
R-PREC	0.3521	0.3566	0.3568	0.3582	**0.3598**	0.3576	0.3597	**0.3598**	0.3587	0.3567	0.3543	0.3418	0.3583

3 Results Discussion

Our experiments describe in several measures. The standard *TrecEval* tool which is provided by *NIST* is used for evaluation [23]. Table 2 and Fig. 2 depict well-known interpolated precision-recall diagram for Hamshahri corpus[1] and Table 3 is our submitted running on CLEF test set. In web retrieval tasks, the number of terms in a query is usually small like Hamshahri queries. If the terms cannot provide enough information of the user's need, the retrieval result may be poor. These are known as weak queries [13]. The TREC Robust track [13] was created in 2003 to focus on poor performing queries. Several new measures were introduced to evaluate the effectiveness on weak queries. Since 2004, another new measure Geometric MAP (GMAP) [19] was introduced as an alternative to the mean average precision (MAP). GMAP takes the geometric mean of average precisions of all the queries instead of their arithmetic mean. Table 3 shows a comparison between best initial results [3], [5], [14] and LCA approach using different variants, on Hamshahri corpus.

In this paper we evaluate two different variants of K-means, Spherical k-means [15] and Euclidean K-means [4]. As you see in Table 2 and Table 3, Euclidean K-means give the better results than Spherical K-means between all measures.

Although Euclidean k-means appears to give the better results between all variants and all measures (except PDDP_2MEANS) especially PDDP, we note that these plots report mean values attained by k-means and related variants. In practice, a single run of k-means may lead to poor results. As a result, a "good" partitioning may require several executions of the algorithm.

Compared to the basic algorithm, PDDP_2MEANS [10] appears to give the best results between all variants and all measures, even better than k-means.

As you see in Table 2, Table 3 and Fig. 2 LCA make valuable improvement against initial retrieval on Hamshahri corpus. Regarding CLEF train set, we get 26% improvement over MAP measure that compatible with same work on Hamshahri corpus (Table 2, Table 3 and Fig. 2) but we have some problems with test set. As you see in Table 4 our results are weak. In some cases, LCA reduce initial result that we have never such cases on Hamshahri corpus. It must be a program bug or something else; by the way we're working on it now.

4 Conclusion

In LCA approach, the context of a document is considered in the retrieved results by the combination of information search and local cluster analysis, cause first: relevant cluster tailored to the user information need and improve the search results efficiently, second: make high-precision system that contain more relevant documents at top of the result list [1]. Clustering algorithm is a key factor in LCA. We evaluate two well-known clustering algorithms (PDDP and K-means) plus some state of the art approach to improve the weakness of both and create superior results. As you see in section II, we introduce a variant of PDDP (PDDP_2MEANS) that have better results than Euclidean K-means without some shortcomings of k-means.

References

[1] Amir Hossein Jadidinejad, Hadi Amiri, "Local Cluster Analysis as a Basis for High-Precision Information Retrieval", *In Proceeding of INFOS2008 International Conference on Informatics and Systems*, Egypt, 2008.

[2] A. Aleahmad, H. Amiri, F. Oroumchian, and M. Rahgozar. "Hamshahri: A standard Persian text collection". *White Paper, Database research Group, University of Tehran*, 2008.

[3] A. Aleahmad, P. Hakimian, F. Mahdikhani, and F. Oroumchian. "N-gram and local context analysis for Persian text retrieval". *International Sympo-sium on Signal Processing and Its Applications*, 2007.

[4] A. K. Jain, M. N. Murty, and P. J. Flynn. "Data clustering: a review". *ACM Comput. Surv.*, 31(3):264–323, 1999.

[5] A. Nayyeri and F. Oroumchian. "Fufair: a fuzzy farsi information retrieval system". In *AICCSA '06: Proceedings of the IEEE International Conference on Computer Systems and Applications*, 2006.

[6] A. Tombros, R. Villa, and C. J. V. Rijsbergen. "The effectiveness of query-specific hierarchic clustering in information retrieval". *Inf. Process. Manage.*, 38(4):559–582, 2002.

[7] D. Boley, "Principal direction divisive partitioning", *Data Mining and Knowledge Discovery*, 2(4):325–344, 1998.

[1] http://ece.ut.ac.ir/dbrg/hamshahri/

[8] D. Boley, "A scalable hierarchical algorithm for unsupervised clustering", *Data Mining for Scientific and Engineering Applications*, Kluwer Academic Publishers, Norwell, MA, 2001.
[9] D. Littau and D. Boley, "Clustering very large datasets with PDDP", *Grouping Multidimensional Data: Recent Advances in Clustering*, Springer, New York, pp. 99–126., 2006.
[10] D. Zeimpekis and E. Gallopoulos, "Principal Direction Divisive Partitioning with Kernels and k-Means Steering", *In Survey of Text Mining II: Clustering, Classification and Retrieval*, Michael W. Berry and Malu Castellanos eds., pp. 45-64, Springer, 2008.
[11] D. Zeimpekis and E. Gallopoulos. "TMG: A MATLAB toolbox for generating term-document matrices from text collections", *Grouping Multidimensional Data: Recent Advances in Clustering*, Springer, New York, pp. 187–210, 2006.
[12] P. Ogilvie and J. Callan. "Experiments using the Lemur toolkit". In Proceedings of the 2001 Text REtrieval Conference (TREC 2001) . National Institute of Standards and Technology, special publication 500-250. pp. 103-108, 2002.
[13] E.M. Voorhees. Overview of TREC 2003. In *TREC*, pp. 1–13, 2003.
[14] H. Amiri, A. AleAhmad, F. Oroumchian, C. Lucas, and M. Rahgozar. "Using owa fuzzy operator to merge retrieval system results". *The Second Workshop on Computational Approaches to Arabic Script-based Languages,* LSA 2007 Linguistic Institute, Stanford University, USA, 2007.
[15] I. S. Dhillon and D. S. Modha, "Concept decompositions for large sparse text data using clustering", *Machine Learning 42*, *no. 1*, pp. 143-175, 2001.
[16] J. Kogan, *Introduction to Clustering Large and High-Dimensional Data*, Cambridge University Press, New York, 2007.
[17] J. Xu and W. B. Croft. "Query expansion using local and global document analysis". In *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 4–11, New York, NY, USA, 1996.
[18] J. Xu and W. B. Croft. "Improving the effectiveness of information retrieval with local context analysis". *ACM Trans. Inf. Syst.*, 18(1):79–112, 2000.
[19] S.E. Robertson, "On GMAP: and other transformations", *In CIKM*, pp. 78–83, 2006.
[20] S. M. Savaresi and D. L. Boley. "A comparative analysis on the bisecting k-means and the pddp clustering algorithms". *Intell. Data Anal.*, 8(4):345–362, 2004.
[21] Hadi Amiri, Hosein Hojjat, Farhad Oroumchian. Investigation on a Feasible Corpus for Persian POS Tagging. 12th international CSI computer conference, Iran, 2007.
[22] Farhad Oroumchian, Samira Tasharofi, Hadi Amiri, Hossein Hojjat, Fahime Raja. Creating a Feasible Corpus for Persian POS Tagging. Technical Report, no. TR3/06, University of Wollongong in Dubai, 2006.
[23] National Institution of Standards and Technology: http://trec.nist.gov/trec_eval/