Sundanese ancient manuscripts search engine using probability approach

Mira Suryani, Setiawan Hadi, Erick Paulus, Intan Nurma Yulita, and Asep K. Supriatna
University of Padjadjaran, Jalan Raya Bandung Sumedang KM21 Jatinangot 45363, INDONESIA
E-mail: mira.suryani@unpad.ac.id, setiawan.hadi@unpad.ac.id, erick.paulus@unpad.ac.id, intan.nurma@unpad.ac.id, a.k.supriatna@unpad.ac.id

Abstract. Today, Information and Communication Technology (ICT) has become a regular thing for every aspect of live including cultural and heritage aspect. Sundanese ancient manuscripts as Sundanese heritage are in damage condition and also the information that containing on it. So in order to preserve the information in Sundanese ancient manuscripts and make them easier to search, a search engine has been developed. The search engine must has good computing ability. In order to get the best computation in developed search engine, three types of probabilistic approaches: Bayesian Networks Model, Divergence from Randomness with PL2 distribution, and DFR-PL2F as derivative form DFR-PL2 have been compared in this study. The three probabilistic approaches supported by index of documents and three different weighting methods: term occurrence, term frequency, and TF-IDF. The experiment involved 12 Sundanese ancient manuscripts. From 12 manuscripts there are 474 distinct terms. The developed search engine tested by 50 random queries for three types of query. The experiment results showed that for the single query and multiple query, the best searching performance given by the combination of PL2F approach and TF-IDF weighting method. The performance has been evaluated using average time responds with value about 0.08 second and Mean Average Precision (MAP) about 0.33.

1. Research Methods
There are two major phases in this study: build the collection and build Sundanese ancient retrieval system. The two activities described in the following section.

1.1. Build The Collection
The sub-phase of build the collection consists of three steps: data acquisition, transliteration, and segmentation including annotation. The detail of four steps describe below.

1.1.1. Data Acquisition
The acquisition data of Sundanese ancient manuscript located in Situs Kabuyutan Ciburny, Garut, West Java. Because of the documents are fragile, so, we cannot hold it for a long time. The image of the documents should be acquired to make it easier to be processed in the next step.
1.1.2. Transliteration  After the image of the Sundanese ancient manuscript obtained, then the transliteration of the documents done by the philologist. There are three philologists read the image of Sundanese ancient manuscript, then re-write the content of manuscript in Latin letters. The results of this step are the Sundanese documents that contain the content of the Sundanese ancient manuscript.

1.1.3. Segmentation and Annotation  Segmentation and annotation step done by collaboration between the researchers and the philologists. The philologists segmented and annotated the documents that obtained from the transliteration step by word manually. The obtained results, then convert to digital segmentation and annotation by the researchers using Aletheia [2]. The results of this step are segmented words, annotation in Latin letters, and the coordinates of words in the documents.

1.2. The Development of The Sundanese Ancient Manuscript Search Engine

![Figure 1](image)

Figure 1. The development of Sundanese ancient manuscript search engine

Figure 1 shows the framework of Sundanese ancient manuscript search engine. The collection of Sundanese ancient manuscripts in Latin version processed in weight terms phase. There are three methods of weighting terms that used in this study: term occurrence (to), term frequency (tf), and term frequency â€“ inverse document frequency (tf-idf). After calculating the term weights, then the set of query runs in three different probabilistic methods. The three probabilistic method set with different parameter values that give based performance based best practices. The parameter values of three probabilistic methods can be seen in experiment and results part. The results of search deliver in interactive interface and the mean average precision of three methods compared.

1.3. Bayesian Network Model

Based on the reference [3], BNM is described as a directed acyclic graph (DAGs) where the nodes represent random variables. The random variables that use in this study are the documents, the terms, and inference results. The arcs in graph expressed causal relationships between variables called the queries. The conditional probabilities described the strength of this dependency relationship. The BNM is one of the probabilistic models that provide clean formalism that combined multiple sources of evidence from a document. Sources of evidence used is past queries, feedback past cycles, and distinct query formulations. For every result search that given by the system according to the query will be saved and used again in the next search process which involved the resemblance query.
Let $t$ be a $y$-dimensional vector defined by $t = (t_1, t_2, ..., t_y)$ where $t_1, t_2, ..., t_y$ are binary random variables $t_i \in \{0, 1\}$. These variables define the $2^y$ possible state for $t$. Let $d_j$ be a binary random variable associated with a document in the collection and let $q$ be a binary random variable associated with the user query. The inference can be calculated by:

$$P(q \wedge d_j) = \sum_{\forall k} P(q|t) \cdot P(d_j)$$

$$(\prod_{(q_i|on(t_i))=1} \cdot P(t_i|d_j) \cdot \prod_{(q_i|on(t_i))=0} \cdot P(t_i|d_j))$$

$$P(q \wedge d_j) = 1 - P(q \wedge d_j)$$

where $P(t_i|d_j) = 1 - P(t_i|d_j)$

The $P(q|t)$ is term-query beliefs that can be obtained from (2). A reasonable default value for $w_q = 1$.

$$P(q|t) = \sum_{k} f_{i,j} \cdot w_q$$

$$P(q|t) = 1 - P(q|t)$$

The $P(d_j)$ is prior probability of observing a document $d_j$ that can be set $\frac{1}{N}$ where $N$ is the total number of documents in the collection.

$$For P(d_j) = 1 - \frac{1}{N}$$

The prior probability of the document-term beliefs can be obtained from (4) below.

$$P(t_i|d_j) = \alpha + (1 - \alpha) \cdot f_{i,j} \cdot \frac{IDF_i}{IDF}$$

$$P(t_i|d_j) = 1 - P(t_i|d_j)$$

Where $f_{i,j}$ is normalized term frequency given by $f_{i,j} = \frac{k_{i,j}}{\max f_{i,j}}$ and $\frac{IDF_i}{IDF}$ is inverse document frequency variable given by (5):

$$\frac{IDF_i}{IDF} = \frac{\log \frac{N}{n_i}}{\log N}$$

Based on empirical evidence, the value of $\alpha$ is 0.4. The final step is combining the evidence. It can be done by retrieve all the type of the past query that relevance to the new one. In order to get the ranking provided by inference network, equation (6) should be calculated.

$$P(I \wedge d_j) = \sum_{l} P(I|\bar{t}) \cdot P(\bar{t}|d_j) \cdot P(d_j))$$

$$= \sum_{l} P(1 - P(\bar{q}|\bar{t})) \cdot P(\bar{q} | \bar{t}) \cdot P(\bar{t}|d_j) \cdot P(d_j))$$

1.4. Divergence From Randomness with PL2 Distribution

Based on the reference [4], the idea of Divergence From Randomness (DFR) method is to compute the term weights by measuring the divergences between a term distribution proceeds by the random process and the actual term distributions. A word that contains a bit of information is generally scattered in the collection. We call $C$ as a collection, then a term $k_i$ has a probability distribution with respect to $C$ as follows $P(t_i|C)$, then the probability of a number of other information is $- \log P(t_i|C)$. Term distribution in which a term referred to as an elite set, then the probability to a document given by $P(t_i|d_j)$. The smaller probability of occurrence of the $k_i$
term in the document indicates increasingly rare and important that term in representing the document.

There are many probability distribution families that can be implemented in DFR Framework. One of the distribution that can be implemented is PL2. Based on [5], the indexing weight of the query terms $w_{ij}$ in the document $d_j$ calculated with (7):

$$w_{ij} = -\log_2(\text{Prob}_{ij}^1(tf)) \cdot (1 - \text{Prob}_{ij}^2)$$ (7)

For $\text{Prob}_{ij}^1$, can be replaced by PL2 distribution. PL2 is Poisson model with Laplace after-effect and normalisation 2. The formula of PL2 distribution can be seen in (8).

$$\text{Prob}_{ij}^1 = \frac{e^{-\lambda_j} \cdot \lambda_j^{f_{ij}}}{f_{ij}!} \text{where } \lambda_j = \frac{to_j}{n}$$ (8)

The $to_j$ indicates the number of occurrences of term $t_i$ in the collection. The $\text{Prob}_{ij}^2$ can be replaced by normalizing term frequency that can be calculated with the following equation (see Eq (9)).

$$\text{Prob}_{ij}^2 = \frac{tf_{nij}}{tf_{nij} + 1}, \text{with}$$

$$tf_{nij} = tf_{ij} \cdot \log_2[1 + \frac{\text{meandl}}{l_i}]$$ (9)

The $c$ is a constant with value 1.5, the $\text{meandl}$ is the average document length, the $n$ = number of documents in the corpus, and $l_i$ = the length (number indexing terms) of document $d_j$ in the collection.

1.5. Divergence From Randomness with PL2F Distribution

Another probability distribution in DRF adopted from Craig, Christina, and Iadh in [1] is PL2F. PL2F is derivative from PL2. The relevance of weighting score of a document $d_j$ for a query terms is given by (10):

$$\text{score}(d_j, q) = \sum_{t \in q} qt w \cdot \frac{1}{tf_{nij} + 1}(tf n \cdot \log_2 tf n + (\lambda - tf n) \cdot \log_2 e + 0.5 \cdot \log_2(2\phi \cdot tf n))$$ (10)

where $\lambda$ given by $\frac{tf}{n}$, $F$ is the frequency of the query term $t$ in the collection, and $n$ is the number of document in whole collection. The query term weight $qt w$ is given by $\frac{qt f}{qt f_{\text{max}}}$, where $qt f$ is frequency of query term and $qt f_{\text{max}}$ is the maximum frequency of query term among the query terms. The $tf n$ corresponds to the normalized term frequency which is given by (11):

$$tf n = \sum_f (w_f \cdot tf f \cdot \log_2(1 + c_f \cdot \frac{\text{avg} f}{l_f}))$$ (11)

where, $tf f$ is the frequency of the term $t$ in the field $f$ of the document $d_j$; $\text{avg} f$ is the average length of whole documents in the collection; and $l_f$ is the length in tokens of field $f$ of the document $d_j$. The $w_f$ and $c_f$ are the hyper-parameter that controlled the contribution of watch field. Based on the condition of Sundanese ancient manuscript in this study, the field that use in both parameter are the body of the document with values $c = 4.10$ and $w = 1$. 
Table 1. The Collection Statistics

| View Point                  | Value  |
|-----------------------------|--------|
| Size                        | 100 MB |
| Number of Documents         | 12     |
| Number of Terms             | 823    |
| Number of Distinct Terms    | 474    |
| Mean of docs                | 68.58  |
| St Dev.                     | 15.56  |
| Median                      | 73.5   |
| Max                         | 87     |
| Min                         | 29     |
| Topics                      | Ramayana story |

2. Experiments and Results

The collection of Sundanese terms obtained from Sundanese ancient manuscript describe in Table 1.

When the weighting term process done, each terms has their own weight, then the set of query randomly picked to run in the three probabilistic approaches. The parameter values for three different probabilistic methods shown in Table 2.

Table 2. The Parameter Settings

| Method        | Parameter | Value |
|---------------|-----------|-------|
| BNM           | $w$       | 1.00  |
|               | $\alpha$  | 0.40  |
| DFR-PL2       | $c$       | 1.50  |
|               | $\text{meandl}$ | 68.58 |
| DFR-PL2F      | $c$       | 4.10  |
|               | $w$       | 1.00  |
|               | $\text{meandl}$ | 68.58 |

The set of queries in the proposed system. The results of search shown in interactive interface on the browser. The example of search results can be seen in Figure 2.

Figure 2. Beta version of Sundanese ancient manuscript search engine

When the proposed system tested by 50 queries, the query latency can be calculated. The different average time responds from three probabilistic shown that the best average time
responds given by DFR-PL2F and the lowest respond given by BNM. The ratio of query latency of three methods is 16:1:1 respectively to BNM, PL2, and PL2F. The complete results of average time respond that indicates the query latency can be seen in Table 3.

| Method     | Average Time Response (sec) |
|------------|-----------------------------|
| BNM        | 1.30 s                      |
| DFR-PL2    | 0.09 s                      |
| DFR-PL2F   | 0.08 s                      |

Table 3. The Average Time Responses of Query Search

Based on an experiment, there is no difference of mean average precision for the three probabilistic methods. The 50 queries that run in the system gave the MAP value about 0.34. The difference of mean average precision can be seen when the proposed system tested by two terms of query type. Even though three probabilistic methods have a derivation of MAP value, but the best performance given by DFR-PL2F with 0.22 for 50 queries. Figure 3 shows the MAP value for the two terms query that applied in the proposed systems.

Figure 3. The comparison of mean average precision for two terms query

When the proposed system tested by three terms query type, there are significantly different with the MAP value. The best performance given by the DFR-PL2F method with MAP value
is about 0.33 and the lowest performance given by the BNM method with the MAP value is about 0.25. If we looked at precisely at the Figure 3 and 4, even though the BNM method gives the lowest performance, but there are escalation results of the MAP value when tested from 10 queries until 50 queries per batch.

3. Conclusions and Future Works
Based on the experiment results, the best query latency performed when using the PL2F method with average time 0.08 second. The MAP evaluation shown that the best performance of the proposed system is given by the PL2F method, especially for 2 and 3 terms query. Even though the PL2F gives the best performance in the latest sum of the collection, the BNM method shown the better progress value of the MAP than the PL2 and the PL2F method but still the lowest. There are many recommendations that should be done in the future, such as create stop word list in Sundanese language, adding the collection, apply the variance of Sundanese stemming methods, and develop other probability methods for results search improvement.

Acknowledgement
The authors would like to thank to DRPMI Universitas Padjadjaran, the Faculty of Nat. Science and Mathematics, the Minsitry of Research, Technology, and Higher Education that gives The International Collaboration and Publication grant in 2016, the philologist from Sundanese Centre Studies of Universitas Padjadjaran, the Situs Kabuyutan Ciburuy Garut, and the L3i Lab of University La Rochelle France as a collaboration partner.

References
[1] Craig M, Crishtina L, and Iadh O 2007 Terrier Takes on The Non-English Web in SIGIR âĂŹ07 iNEWS07 Workshop, Amsterdam
[2] Clausner C, Pietschacher S, and Antonacopoulos A 2011 Aletheia - An Advanced Document Layout and Text Ground-Truthing System for Production Environments in International Conference on Document Analysis and Recognition, Beijing
[3] Omri M N, Amira K, and Kamel G, 2012 A New Information Retrieval Model Based on Possibilistic Bayesian Networks in International Conference on Computer Related Knowledge, Sousse
[4] Ricardo B Y and Berthie R N 2011 Modern Information Retrieval London: Pearson Addison Wesley
[5] Jacques S, Ljiljana D, and Mitra A 2011 Information Retrieval with Hindi, Bengali, and Marathi Language: Evaluation and Analysis in Third International Workshop FIRE, Bombay.