Evaluation of probabilistic models for word frequency and information retrieval

Fateme sharifi jahantigh, Fahime zarei

Abstract— **Finding data recovery data (usually documents) without the (usually Text) in terms of user needs through a large set usually stored on computers is described. In this study, the frequency distribution. of word associations and their potential for Information Retrieval (IR) the model and Huge frequency the word in the text of the review. Our on probabilistic models IR and will be the focus of their basic features.**

Index Terms— **Word frequency, probability models, language models, IR, ranking.**

I. Introduction

The volume of information available on the computer, users to access and search resources efficient methods need. Data recovery organizing unstructured data modeling available on the database takes place access of users through the massive collection of documents / the information accelerate. Naturally, IR as a sub-domain processing Natural language (NLP) began to emerge. Information request form when the user performs a query language as an indexed document to be displayed. A function to apply Reference to the document presented to the user adjusts the ranking list.

II. Main Content

An information retrieval system Includes three elements:

- A query model
- A Document Model
- A Function To Name The rate of recovery scenarios (RSV), query matching documents will be shown in Figure 1. Much larger than the the documents provide a better response to requests by default. Most of the first models IR morphs as First-order logic are considered. From this point, consider a document d if it Included Q request is based on logical rules.

(1) $$RSV(q,d)=\begin{cases} 1 & if\ d => q \\ 0 & otherwise \end{cases}$$

Manuscript received September 09, 2014.

Fateme sharifi jahantigh, *Department of information technology*, Iran
Fahime zarei, Department of information technology / Young Researchers and Elite Club, Kerman Branch, Islamic Azad University, Kerman, Iran

Vector models Howe requested documents in Euclidean spaces provided. Euclidean spaces of any dimension indexing are a word or phrase. After that, the degree of similarity between a document d and query q is the angle between the two vectors can be calculated.

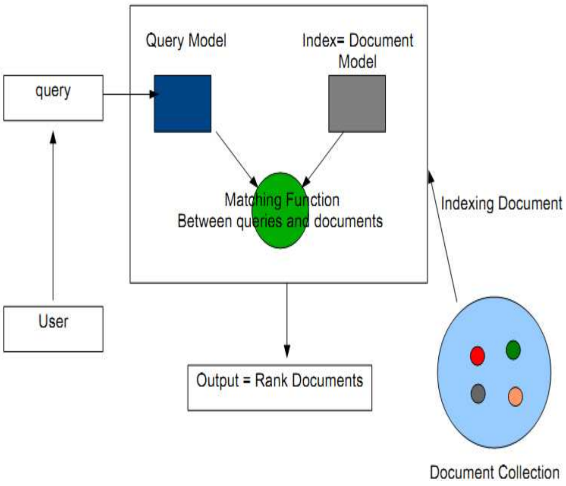

Fig 1: Information Retrieval System Architecture [1]

(2) $$RSV(q,d) = \cos(\vec{q}, \vec{d})$$

Finally, probabilistic models of information retrieval queries and documents adopted as a result of random processes. Many of the problems IR A probabilistic framework is dissolved. One way, all probabilistic models IR Given that the words and their frequency in a document or set of documents can be considered as random variables. Therefore, it may be observed that the frequency of the word as random events [1]. Hence, probabilistic models of the selected probability distribution of the model, documents and queries are supported. Probabilistic models can be identified by three elements the probability distribution of the documents modeled Pdoc, the probability distribution application modeled Pquery and the function H the distribution edit with matches.

random variable D modeling document d, Q random variable modeling query q, then a Model the IR can be defined as follows:

(3) $$D \sim P_{doc}\ (0 \mid \lambda)$$
$$Q \sim P_{query}\ (0 \mid \theta)$$
$$RSV(q,d) = H(P_{query}(Q = q \mid \theta)\ , P_{doc}(D = d \mid \lambda))$$

III. Helpful Hints

A. Figures and Tables

burstiness by Katz defined as follows: the multiple events from a single word or phrase in a document with the fact that many other documents containing any of these words or phrases are not in conflict. [1-2]. the word "burstiness describe the behavior words that tends to They appear, when they appear in a document, it will be much more likely Appear again. Polynomial model it is a very popular model. The first classification model used a simple Bayes [3]. Then the IR With the so-called model Language was called. Polynomial distribution One Generalized binomial distribution Several Variable. Polynomial model during a document l_d And the θ encryption than any words, and words are assumed to be twist in independent of each other. Seen a document is simply a sign of the bag in which the words are independent of each other. This means that the event statistically independent of the word for example, a document may include the words: (soviet; president; US; soviet; cold; war) is Thus, the occurrence US And soviet are independent of each other. So soviet event repetitive. Shown in Figure 2. Random variables X_w Binomial distribution shows that the mean and variance.

(4) $E(X_w) = L_d \theta_w$

$Var (X_{w)} = 1_d \theta_w (1-\theta_{w)})$

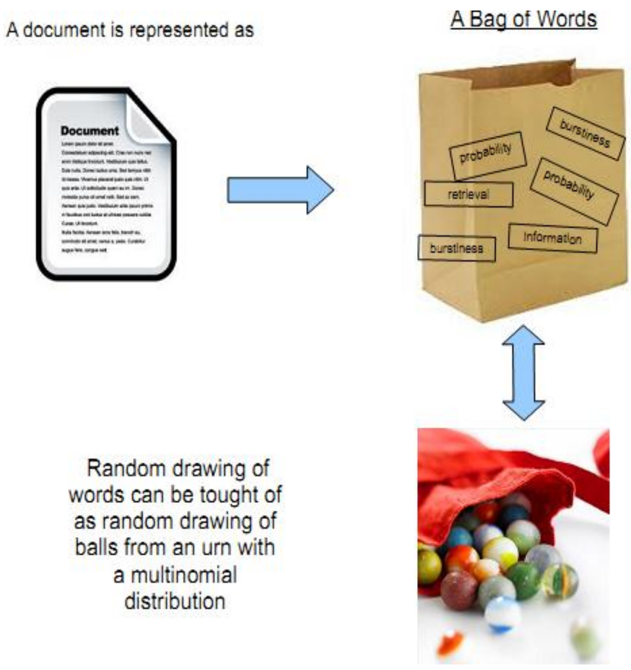
A document is represented as
A Bag of Words
Document
probability
burstiness
retrieval
probability
burstiness
Information
Random drawing of words can be tought of as random drawing of balls from an urn with a multinomial distribution

Fig2: A package similar words with bags of balls in polynomial model [1]

Thus the variance of the distribution is controlled by the average of the limits Model. Moreover, the distribution of polynomials. It is very appropriate because it estimates It's simple. Maximum likelihood estimates if **θ m.le are:**

(5) $$w^{\hat{}}\theta = \frac{Xwd}{\sum_w Xwd}$$

All systems of information retrieval include a model query, a model document and a function for the implementation query with documents. These three components are required for an IR engine. Despite the fact that the machine learning approach IR recently, IR has been a great success.

We focus first ad hoc retrieval. Since the generating function and differential approaches the scenario is similar to the ad hoc [4]. Principle rankings Probabilistic, Language models, and the divergence from randomness, three of family recovery with its special features are rely on the word distribution. Okapi, for example, it is assumed that the frequency of the word followed is combined by two Poisson distributions. Framework divergence of randomness (DFR) by Amati and van Rijsbergen [5] suggested that the number of distribution The use of the geometric distribution, Poisson distribution and laplas Succession law has an important role to play. A brief description of the IR models Explains.

2-1 ranks the probability ranking principle (PRP): This model assumes that one class of relevant documents and a class of documents Irrelevant to a query There. The idea of ranking relevant documents likely to be estimated It is. Models featured in this family Okapi or BM25 named.

2-2 language models (LM): The basic idea of the language model to estimate the probability a request that is derived from documents model P (q | d). nowadays Language models are very popular

2-3 Divergence randomly (DFR) This model tries to quantify the importance of a word in a document collection show in action. Thus, the weight of a word in document can be a function Shannon information measure based on the ranking of all models of The probability [6], [1] Assuming that relationship with a document or a request can be encoded as a random variable. We will focus R_q Random variable connection request q. This assumes in 70 decade that developed a significant impact on information retrieval model has had. If a reference retrieval system response to each request is a ranking of the documents in descending order of probability of being linked to request a place the presentation of data. Effectiveness of the overall system is better. Rankings the likelihood of a direct result of the base of making Bayes. Suppose you have:

(6)

$$P(\text{error} \mid X^d) = \begin{cases} (P(R = 1|X^d) \; \textit{if one chooses } R = 0 \\ (P(R = 0|X^d) \; \textit{if one chooses } R = 1 \end{cases}$$

Then, if a choice R = 0 (non-relevant document) are P (R = 1 $|X^d$)> | P (R = 0$|X^d|$ This decision led to Larger error is a false choice. Therefore, to choose the assumption that the maximum (P (R$|X^d$) *and* the probability of error at least bring enough. Assumption that the documents are (statistically) independent of the law of the place documents with their associated probability is reduced. Figure 3 shows the IR approach is likely the rankings.

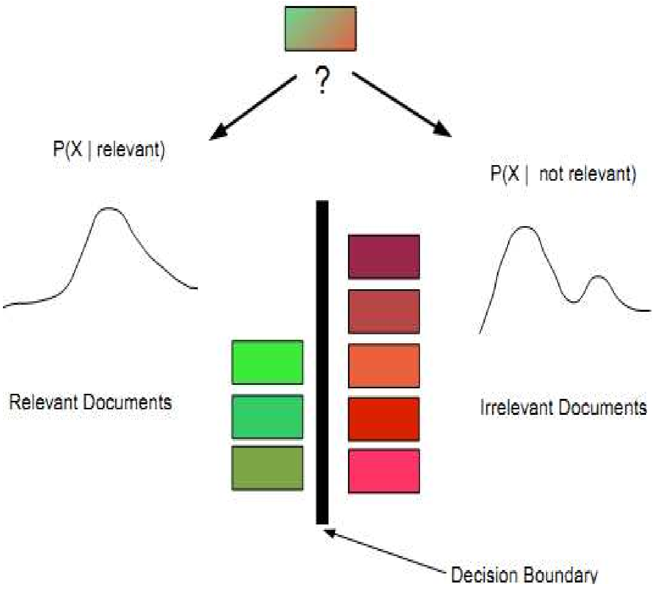

Fig 3: The probability ranking principle [1].

However, it can predict the likelihood of your Test and test sequentially correct, improve their estimates. However, when your chance to This principle cannot be correctly graded suboptimal [1], [7] showed., With the assumption that some relevant, non-relevant documents known, a probabilistic model is assumed to be The probability that a new document is relevant or not estimated. After the first phase of recovery, users can to explain Documents and probability of relevant that it is updated. So

(7) $P\left(X^{d} \middle| R = r\right)$

As a likelihood product and of function as the sum of points as words common between the query and Considered documents. Example, Robertson this function so that documents are empty The zero points.

(8) RSV(q,d)=$\sum_{w \in q \cap d} \log\left(\frac{P(X_w^d = x | R=1)\, P(X_w^d = 0 | R=0)}{P(X_w^d = 0 | R=1)\, P(X_w^d = x | R=0)} \right.$

Binary independent models (BIR) in this model is assumed to be binary word weighted in documents and query to be binary the probability that the word will appear in a document (A_w) represent the word with binary variable is described. Every word conditionally Together independent in the set of R is given. . $X^d = (10100 \cdot\cdot 1000)$ Two words $W_{1,}$ $W_{2,}$ then y | R = r) = x | R = r) P(X_{w2}^d y | R= r) =P(X_{w1}^d = = x,X_{w2}^d P(X_{w1}^d

Note that a_w = P (A_w = 1 | R = 1), possibly the word w in the related document appears (A_w = 1 | R = 0)

possible word w appears in non-Related documents

P (X^d = ($x_{1,}$ ..., $x_{M)}$ | R = 1) = $\prod_w aw^{xw} (1 - a_w)^{(1-Xw)}$
$P($ X^d = (X_1, , ... , x_M) | R = 0) =
$\prod_w bw^{xw} (1 - b_w)^{(1-Xw)}$

In other words, the documents based on rules have been modeled by independent Bernoulli. Fact that a document is relevant or not relevant. This probability measure various parameters (a_w or b_w) Are described. Based on these assumptions, PRP Be expressed as follows:

(10) RSV (q, d) = $\sum_{w \in q \cap d} \log\left(\frac{aw}{1-aw} \frac{1-bw}{bw} \right)$

The estimated probabilities a_w and b_w Is performed with an iterative process.

1. Defined Initial values , For example = $a_w^0 , b_w^0 = \frac{Nw}{N}$ =0.5*0*

2. A recovery phase with the current parameter values are performed.

3. Parameters are updated If V the number of relevant documents In this stage

V_W number of the document contains the word W, and re-estimates the parameters will be $b_w = \frac{Nw - Vw}{N-V}$ $a_w = \frac{Vm}{V}$

The advantages of use this model is theoretically Well documented and linked to a clear concept of the show. Moreover, the data recovery process to form a portion repetitive That involves the user. However, this model is quite sensitive to initial values Its major drawback presented binary the occurrence of words in document that system performance lowers

Okapi/BM25

BM25 Model Some of the shortcomings of the BIR Will be examined. Originally, BM25 assumes that the frequency of words according to the combination of two Poisson In addition, given that the Collections relevant (R = 1), the Poisson distribution gives more weight Elite component of the non-relevant class Typically, these assumptions result in:

Xd = x | R = 0 ~ 2poisson (β, λE, $\lambda_{G)}$ α> β

Xd = x | R = 1 ~ 2poisson (α, λE, $\lambda_{G)}$

Recall That λE > λ_G Formula PRP with Robertson:

(12) RSV(q,d)=$\sum_{w \in q \cap d} \log\left(\frac{P(X_w^d = x | R=1)\, P(X_w^d = 0 | R=0)}{P(X_w^d = 0 | R=1)\, P(X_w^d = x | R=0)} \right.$

Adding the assumptions in 2Poisson model:

Knowing α> β, we can show that this function means a function of word frequency increases X_w Moreover, limit h, when X_w Will tend to infinity. Values are: (13)

$$\lim_{x \xrightarrow[+\infty]{}} h(x) = \log\left(\frac{\alpha}{\beta} \frac{(\beta e^{-\lambda E + \lambda G} + 1 - \beta)}{(\alpha e^{-\lambda E + \lambda G} + 1 - \alpha)} \right) \approx \log\left(\frac{\alpha}{\beta} \frac{1-\beta)}{1-\alpha)} \right)$$

Approximate this limit using the fact that $\lambda_G < \lambda_E$ Robertson and Walker's idea was to find a function which has the same features as the h is.

Initially, he suggested using a function of the type r(X) = $\frac{X}{X+K}$ *which is rising* tends *toward 1.*

Then, he proposed a multiple of the weight of this last function that models BIR Gives the same level of approximation of the function h.

$$\log\left(\frac{P(X_w^d=1|R=1)\,P(X_w^d=0|R=0)}{P(X_w^d=0|R=1)\,P(X_w^d=1|R=0)}\right) \quad x_w) = \frac{xw}{xw+K}(h^* \qquad (14)$$

$$h^*(x_w) = \frac{Xw}{Xw+K}\log\left(\frac{aw}{1-aw}\ \frac{1-bw}{bw}\right) aw$$

bw , several can be estimated**.**

It is necessary that during the age the document frequency is calculated so instead of using a function of the type

$$(15)\ \frac{X}{X+K}$$

can be selected Function form below

$$.\ \frac{(K1+1)Xwd}{K1\left((1-b)+b\frac{ld}{avgl}\right)+Xwd}$$

Where avgl The median length of the document and the document d in the collection. K $_1$ The default values of 1.2 and b is set to 0.75. Words Normalized frequency query as follows :

By default, K $_3$ = 1000

Finally, $_{\text{consistent}}$ with the initial default $_{\text{values.}}$ (16)

RSV (q, d)

$$= \sum_{w\in Q}\frac{(K3+1)qw}{K3+qw}\ \frac{(K1+1)Xwd}{K1\left((1-b)+b\frac{ld}{avgl}\right)+Xwd}\ \log\left(\frac{N-Nw+0.5}{Nw+0.5}\right)$$

BM25 formula is fairly complex and involves three parameters (K1, K3, b) can probably be optimized on a specific data set. In this model 1995 income and a great success at the polls, such as TREC was recognized. Still, as a model Reference should be considered.

3 - Language Model

language models comes from the speech processing as a probability distribution on a sequence of words are defined.

The main idea language model in information retrieval Ranking documents with probability P (d | q) The probability query can be a document that

D Model To come. Hence, most the related documents likely to produce the requested. Similarities with the vector space model [9], [14] Straightforward. Rather than a document by a vector, a document with a probability distribution, rather than calculating the Euclidean distance Contingency - Divergence KL Is calculated. Figure 4 language Modeling Principle for IR Shows. So any document is required to be a dependent language model, the probability distribution. Ponte and Croft

[10-11] The first models of language For IR r Which later improved in many ways [12]. Zha et al a good overview of the modeling approach Language Offer Reported [13]. Many of these models makes the selection of the polynomial distribution A In the model, the documents said. One of the basic assumptions of language modeling approach is to each document is a document language model, (17), ie, θ^{d} such that

$$P\,(X^d = (x_{wd})\ |\ \theta_d, l_d) = \frac{ld!}{\prod_{i=1}^{n} xwd!}\prod_{w\in d}(\theta wd)^{xwd}$$

estimated document language modelFor each document Doing this, the maximum likelihood estimate of the observer (mle) is often used:

$$(18)\ \hat{\theta}_{wd}^{mle}\ ) = \frac{xwd}{\sum_w xwd}\ \frac{xwd}{ld}$$

$$RSV(q,d) = \log P(q\,|\,\widehat{\theta d}\ ,l_q) = \sum_{w\in q} qw\ \log(\widehat{\theta}\ wd)\ + h(q$$

However, the m.l.e a big problem: if a word appears in a document, this query is assigned probability zero document Given, so that the log likelihood is undefined. To overcome these problems, the soft way to add some Background knowledge on the document language model is used.

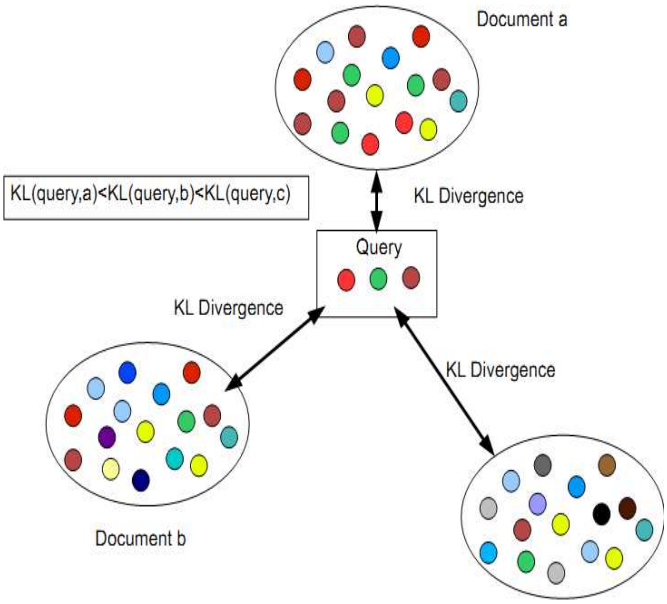

Fig4: a language model approach to information retrieval [1]

4-Soft methods

Dirichlet smoothing and Jelinek-Mercer smoothing The most popular methods smoothing Said. Sets of documents can also be shown with a language model. C Set of documents, then the language model (Multi) is as follows:

(19)

$$\beta_w = P(X_w = 1\,|\,C) = \frac{\sum_d xwd}{\sum ld} = \frac{fw}{f}$$

$$\theta_{wd} = \alpha\ \theta wd^{\hat{}\,mle} + (1 - \alpha))\ \beta w$$

$$P(w\ |\,d) = \alpha\ P^{mle}\ (w\ |\,d) + (1-\alpha)\ P\ (w\,|\,C)$$

Therefore, the α is model Parameter. In general, alpha after some adjustment to maximize your performance in a given set. Scores a document can be Be decomposed into two parts. Part I with words that belong to both the query and the document deals. Part II with words that do not appear in the document. Hence, the latter can be explained in words

$$(20)\ RSV(q,d)\ = \sum_{w\in q} qw \log(\theta wd)$$

Table 1: probabilistic models information retrieval

probabilistic models of information retrieval	**Feature models**		**Benefit**	**Disadvantage**
probability ranking principle	$\log(\frac{P(X^d=x \mid R=1)}{P(X^d=x \mid R=0)}$	Document ranking descent order highest probability of relevance the request is made by the user $P\ (R = 1 \mid X^d)$	To Revise certain shortcomings. The concept of relevance is discussed. Function query matching and documents boost system performance.	Calculate the probability $P\ (R \mid X^d)$ It is difficult to measure Health carefully size
Language models	Smoothing Methods	Ranking documents by the likelihood .p(q|d) The probability of the query a document d that can be to come d model	There Offer document with a probability distribution Cholera contingency - KL Divergence Is calculated.	The current problem Language model of the document (θwd) For each document.
Divergence of Aztsadfy	$RSV(q.d) = \sum_{w\in q\cap d} qw(1 - prob2(twd))\ Inf1(twd)$		In this model, the Shannon information To measure the importance of a word in the document .	---------------

(21) RSV (q, **d**) =

$$\sum_{w\in q, xwd>0} \mathbf{qw}\ \ \log(1+\frac{xwd}{\mu\beta w}) + \mathbf{lq}\ \log\frac{\mu}{ld+\mu} + h(q)$$

In addition, the ratio $\frac{ld}{\mu}$ analog $\frac{\alpha}{1-\alpha}$ for

Jelinek -Mercer smoothing. α, μ According to some performance measures optimized. Yet

, Zhai [12] In order to estimate μ Optimal amount of leave as a risk Proposed document collections. This way, especially as the nature of the judgment need not Estimates have .

KL retrieval Model

The basic Languages model for IR Including calculation The probability of the query For each document in the collection. This model is query can be considered as an example of a generalized random variable [14]., As per the document Set of a query as an example of a distributed polynomial Considered to be a .

(22) $q|\ \theta_q, l_q$ Multinomial (θ_q, L_q)

Query and documents can be probabilistic distance, KL - divergence To compare :

(23) RSV $_{(q,}$ d) =-KL (θ q, l q)

(24) RSV (q, d) = $\sum_{w\in q} \boldsymbol{\theta wd} \log \boldsymbol{\theta wq}$ +h(q)

KL Recovery Model The probability of a query are ranked with the document model and query distributed polynomial.

(25) RSV(q,d) = rank $\sum_{w\in q} \mathbf{qw} \log \boldsymbol{\theta w}d$

IV. Conclusion

Probability models for word frequencies and Recovery was studied in this paper. Whose purpose was to link probable models. Models BM25 Following the principle rankings Two models of the probability of default Compound Poisson frequencies word. Language models are mainly based on polynomial distribution While the models DFR Including Poisson or geometric distribution for the cases. All these recovery models basic is often the need IR Expand or strap, such as document structure calculation. Therefore, we attempt to define and implement a model IR Relying on distributed the bursty are models PRP Need to select a distribution center related class event and we have no indication that the phenomenon burstiness Still Waiting Class re lated not be relevant. General framework This divergence Accidental Which is close to our needs Looks. Performance of the model for pseudo feedback relations theory is based on our analysis.

Thank

Significantly to the development of every individual person is on its back. It is far master thank Dr mahbobeh shamsi guided me in this research.

REFERENCES

[1] s.clinchant. Probabilistic Models of WordFrequencies and Information Retrieval. version 1 - 1 Mar 2012

[2] Kenneth W. Church and William A. Gale. Poisson mixtures. Natural Language Engineering 1:163-190, 1995.

[3] A. Mccallum and K. Nigam. A comparison of event models for na ï ve classi_cation. In The Fifteenth National Conference on Arti_cial Intelligence (AAAI, 1998.

[4] Ramesh Nallapati. Discriminative models for information retrieval. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '04, pages 64 {71, New York, NY, USA, 2004. ACM.

[5] Gianni Amati and Cornelis Joost Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Trans. Inf. Syst., 20 (4): 357 {389, 2002}.

[6] SE Robertson. The Probability Ranking Principle in IR. Journal of Documentation, 33 (4): 294 {304, 1977.

[7] Michael D. Gordon and Peter Lenk. When is the probability ranking principle suboptimal? JASIS, 43 (1): 1 {14, 1992}.

[8] SE Robertson and S. Walker. Some simple e_ective approximations to the 2 - poisson model for probabilistic weighted retrieval. In SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, pages 232 {241, New York, NY, USA, 1994}. Springer-Verlag New York, Inc.

[9] G. Salton and MJ McGill. Introduction to Modern Information Retrieval. McGraw-Hill, Inc., New York, NY, USA, 1983.

[10] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Modern Information Retrieval. Addison-Wesley Publishing Company, USA, 2nd edition, 2008.

[11] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In SIGIR, pages 275 281. ACM, 1998.

[12] Djoerd Hiemstra, Stephen Robertson, and Hugo Zaragoza. Parsimonious language models for information retrieval. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '04, pages 178 {185, New York, NY, USA, 2004. ACM.

[13] Chengxiang Zhai and John La_erty. 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, pages 334 {342, New York, NY, USA, 2001. ACM.

[14] John La_erty and Chengxiang Zhai. Document language models, query models, and risk minimization for information retrieval. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '01, pages 111 {119, New York, NY, USA, 2001. ACM.

Fateme sharifi jahantigh is M.Sc. student of information technology at University of Qom in Iran. She received her B.Sc. degree in information technology from Tabari University of Babol in Iran in 2012. She has research interests in the fields of: Cloud Computing, ERP, Cloud Computing in agriculture and related domains.

Fahime Zarei is M.Sc. student of information technology at University of Qom in Iran. She received her B.Sc. degree in information technology from Kerman University of Iran in 2010. She has research interests in the fields of: network security, security Cloud Computing, LINUX programming, IR and related domains.

.