Improvements of Katz K Mixture Model

Yinghui Xu† and Kyoji Umemura†

A simpler distribution that fits empirical word distribution about as well as a negative binomial is the Katz K mixture. In the K mixture model, the basic assumption is that the conditional probabilities of repeats for a given word are determined by a constant decay factor that is independent of the number of occurrences which have taken place. However, the probabilities of the repeat occurrences are generally lower than the constant decay factor for the content-bearing words with few occurrences that have taken place. To solve this deficiency of the K mixture model, in-depth exploration of the characteristics of the conditional probabilities of repetitions, decay factors and their influences on modeling term distributions was conducted. Based on the results of this study, it appears that both ends of the distribution can be used to fit models. That is, not only can document frequencies be used when the instances of a word are few, but also tail probabilities (the accumulation of document frequencies). Both document frequencies for few instances of a word and tail probabilities for large instances are often relatively easy to estimate empirically. Therefore, we propose an effective approach for improving the K mixture model, where the decay factor is the combination of two possible decay factors interpolated by a function depending on the number of instances of a word in a document. Results show that the proposed model can generate a statistically significant better estimation of frequencies, especially the frequency estimation for a word with two instances in a document. In addition, it is shown that the advantages of this approach will become more evident in two cases, modeling the term distribution for the frequently used content-bearing word and modeling the term distribution for a corpus with a wide range of document length.

KeyWords: Term Distribution Model, Katz K mixture, decay factor

1 Introduction

Term distribution models are a subject of great general interest and a matter of even greater practical importance. In addition to information retrieval (S. E. Robertson 1980), a good understanding of distribution patterns is useful wherever we want to assess the likelihood of a certain number of occurrences of a specific word in a unit of a text.

Linguistic count data often violates the simplest assumptions of standard probability models like a binomial or Poisson distribution. In particular, the inadequacy of the Poisson dis-
tribution for modeling word frequency is well known, and alternatives have been proposed (Mosteller and Wallace, 1984; Church and Gale, 1995; Slava M. Katz, 1996). In the case of the Poisson, a mixture of Poisson distribution has been a commonly-accepted model for the distribution of content words in text documents, with the two-Poisson model viewed as a meaningfully interpretable approximation of a general model (Robertson S. 1994). But, the underlying mechanism of the Poisson process for generating content words in text is incompatible with empirical data. The observed occurrences are not independent of each other, and multiple instances of content words are observed close to each other far too frequently than it would have been if they, indeed, were the outcome of a Poisson process. Another commonly used robust alternative is the negative binomial distribution (Church and Gale 1995), which has the ability to capture extra-Poisson variations in the data. The negative binomial fits term distribution better than one or two Poisson, but can be hard to work with in practice because it involves the computation of large binomial coefficients.

The Katz K mixture model (Katz 1996) is known to be one of the best term distribution models to capture the linguistic properties of term occurrences. There are two kinds of K mixture models proposed by Katz: the two-parameter model and three-parameter model. In the Katz K mixture, terms with specific frequencies are counted according to his innovative definitions of the three-way classifications of documents which are: among documents containing occurrences of any particular content word, documents with a single instance of that word are assumed to be non-topical documents while documents including multiple instances of that word are assumed to be topical documents; All other documents containing no occurrences of that word are considered unrelated documents. Both Katz models assume that the conditional probabilities of repeats are independent of the number of previously-observed occurrences and approximate them by some constant. Whereas, in contrast to the two-parameter models which does not treat the non-topical and topical documents differently, the three-parameter K mixture model handles the non-topical and topical documents separately, and the probabilities of repeat occurrences of the content word are assumed to be the constant only among topical documents. However, while reviewing the Katz K mixture models, although considerable improvements in the modeling of term distribution are made by the three-parameter Katz K mixture compared with the two-parameter model, there are still comparatively critical misjudgments in the distribution of term frequencies observed in the experiments, especially for a word with two or three instances. After examining the assumptions carefully, we noted that the accidental repeats of non-topically-used content words are overlooked in the Katz three-parameter model since the conditional probabilities of repeats in the topical documents are
treated equally without discrimination. Nonetheless, cases of accidental repeats are possible, particularly for frequently used content words or in a long document. In our view, the accidental repeats of a word in a document will probably occur when there are few instances of the word. In this paper, an investigation on the practical characteristic of the conditional probabilities of repeats is conducted. It was discovered that: for function words, the assumption above is acceptable while for content-bearing words, a word with more occurrences will have a greater chance of being repeated and that chance tends to stabilize with an increase in the instances of the word. At the same time, the decay factor, as the deterministic component of the K mixture model, was also explored. The findings in the experiments indicate it is possible to use both ends of the distribution to fit models. That is, not only can document frequencies be used when instances of a word are few, but also tail probabilities can be employed. Both document frequencies for a few instances of a word and tail probabilities for many instances are often fairly easy to estimate empirically. Therefore, we propose an effective approach for improving the K mixture model, where the decay factor is the combination of two possible decay factors interpolated by a function depending on the number of instances of a word in a document. Results show that the proposed model will generate a statistical-significantly improved estimation of frequencies, especially the frequency estimation for a word with exactly two instances in a document. In addition, it is shown that the advantages of our approach will become more obvious in two cases, modeling the term distribution for the content-bearing word that is used frequently and modeling the term distribution for a corpus with a wide range of document length. Our model deals with only the distribution of single term, and thus is simpler than other researches that treat relationship between plural words, such as, Probabilistic Latent Semantic Indexing (PLSI) (Hoffman 1999), Latent Dirichlet Allocation (LDA) (David M. Blei 2003) and Dirichlet Mixtures (DM) (Mikio Yamamoto and Mishina 2003). Treating the relationship between plural words will make the model more interesting and is quite possible to form the theoretical framework of our proposed model, which will be the research of our future work.

This paper is organized as follows: experiment data is firstly described in Section 2. In Section 3, important notations often used in the paper are defined. The Katz K mixture model and experimental results are introduced in Section 4. In Section 5, investigations on several important issues related to the K mixture model are conducted. Based on findings in Section 5, an improved model for the K mixture is proposed in Section 6. Conclusions are presented in Section 7.
2 Experiment Data

The experimental part of this study involves a large collection, English Gigaword Corpora, of English news text (LDC 2004). There are four distinct international sources of English newswires:

- AFE Agence France Presse English Service
- NYT The New York Times Newswire Service
- APW Associated Press Worldstream English Service
- XIE The Xinhua News Agency English Service

The results reported in the paper are based on the experiments using AFE collections, which contain six years news text. There are about 170,969 keywords and 656,269 documents in the entire corpus. To explore the term distribution model for this large corpus, the frequencies of certain items (words, word pairs, substrings, etc.) were counted. Word-based suffix array (Udi Manber 1990) was used to accumulate the necessary frequency and rates (counts of frequency) information. To avoid some unpredictable influences on the modeling term distributions, the experimental conditions were limited to the following:

- The range of the document length in the AFE collection is quite wide, from 1 to 6,950, and the document length of most of the documents is within the range from 200 to 800. To avoid influences on term distribution, statistical information of the terms was extracted from a subset of the total corpus in which document length were limited from 300 to 500. The number of documents in this subset is 153,132.
- In the experiments, to find the characteristics of word repetitions, words with fat-tail properties were used. A fat-tail properties means that there are two or more document having 10 or more appearances of the word in the test corpus.
- There were many types of words in the test corpus. They were classified into two groups based on their linguistic functionality. Firstly, words such as adjectives, pronouns, conjunctions, prepositions, and auxiliary verbs were chosen and called function words. Their occurrences are independent from what the document is about. Secondly, in contrast to function words, some proper nouns and compound noun terms, intrinsically content entities, are extracted as content-bearing words. An examination of distribution was conducted, respectively, for these two types.
- The counting of terms was case-insensitive, and no stemming treatment was conducted on the corpus.
3 Notations

Before introducing the Katz K Mixture model, some often used symbols which represent empirical statistical values in the collection were defined as follows:

Definitions 1

- $w$: terms (words, word pairs, substrings etc.) that need to be counted.
- $d$: a document containing a sequence of terms.
- $D$: a set of documents, usually called the corpus.
- $N$: $N = |D|$, the size of the corpus.

Definitions 2

- $tf(w, d)$: the number of occurrences of $w$ in $d$
- $cf(w) \equiv \sum_{d \in D} tf(w, d)$
- $df(w) \equiv |\{d | tf(w, d) \geq 1\}|$

These three definitions are commonly used in information retrieval literature. They are usually called within-document term frequency, corpus frequency, and document frequency respectively.

Definitions 3

- $df(k; w) \equiv |\{d | tf(w, d) \geq k, d \in D\}|$

The $df(k; w)$ value, as a kind of generalized document frequency, is the number of documents that contain $w$ for $k$ or more times. In particular, we have: $df(1; w) = df(w)$ and $N = df(0; w)$.

Definitions 4

- $ddf(k; w) \equiv |\{d | tf(w, d) = k, d \in D\}|$

The $ddf(k; w)$ value, as a type of differential document frequency, is the number of documents containing the word $w$ exactly $k$ times. With the above definitions, it is easy to obtain:

\[
    df(k; w) \equiv \sum_{i \geq k} ddf(i; w)
\]

\[
    ddf(k; w) = df(k; w) - df(k + 1; w).
\]

Definitions 5

- $cdf(k; w) \equiv \sum_{i \geq k} df(i; w)$

For expository convenience, the definition of $cdf(k; w)$, as a kind of cumulative document frequency, is introduced in this paper. The generalized document frequency, $df(k; w)$, can be recovered from $cdf(k; w)$. The properties of the cumulative frequency are:

\[
    cdf(1; w) = cf(w)
\]

\[
    df(k; w) = cdf(k; w) - cdf(k + 1; w)
\]
Definitions 6

\[ P_w(k+1|k) = P\left(\frac{tf(w,X) \geq k+1}{tf(w,X) \geq k}\right) \]  

where: \( X \) is the random variable for \( d \)

The \( P_w(k+1|k) \) value, used by Katz (Katz 1996), represents the likelihood of a term having more than \( k \) occurrences in a document given that exactly \( k \) occurrences have already taken place. Through the maximum likelihood estimation (MLE), the value of the conditional probabilities can be estimated empirically as \( \hat{P}_w(k+1|k) = \frac{df(k+1;w)}{df(k;w)} \). Particularly, in the case of \( k = 1 \), the \( \hat{P}_w(2|1) \) is estimated by \( \frac{df(2;w)}{df(w)} \) and this statistic had been used as the approximation of adaptation (Church 2000). Adaptation is a term borrowed from the literature on language modeling for speech recognition (Jelinek 1999).

4 Katz K mixture

The Katz K mixture is a simpler distribution that fits empirical word distributions as well as a negative binomial model (Christopher D. Manning 1999). There are two kinds of K mixture models proposed by Katz, the two-parameter model and a three-parameter model. First, we have introduced the definition of a three-way classification of documents (Abraham Bookstein 1974) as used by Katz in his model descriptions. In the proportion of documents that contain occurrences of any particular content word, documents with only a single instance of that word are assumed to be non-topical documents, while documents including multiple instances of that word are assumed to be topical documents; All other documents containing no occurrences of that word are deemed as unrelated documents. Both Katz models assume that the conditional probabilities of repeats are independent from the number of previously-observed occurrences, and approximate them by some constant. Descriptions of these two models are presented, respectively, in the following sections. Here we need to point out that to simplify the model description, the statistics of the words, \( df, cf, df(k), ddf(k) \), and \( cdf(k) \) are used in what follows without specifying the word \( w \).

4.1 The conventional Katz K mixture - two-parameter model

In the two-parameter model, occurrences in non-topical documents and topical documents are not treated differently. Accordingly, the probabilities of repeats as the probabilities of a word which has exactly \( k \) occurrences in a document is modeled by (Christopher D. Manning 1999):

\[ P_{Katz2}(tf(w,X) = k) = (1 - \alpha) \delta_{k,0} + \frac{\alpha}{\beta + 1} \left( \frac{\beta}{\beta + 1} \right)^k \]

where:  \[ \delta_{k,0} \begin{cases} 1 & \text{iff } k = 0 \\ 0 & \text{otherwise} \end{cases} \]
Table 1 Actual and estimated numbers of documents with $k$ occurrences for six function words and six content-bearing terms. K'1 represents the conventional Katz model and K'2 is the Katz's three parameter model. Samples show that two Katz's K mixture model are both work well for function words. While, for content-bearing words, the Katz's three parameter model performs much better than the conventional model.

The parameter $\alpha$ and $\beta$ are estimated using the observed mean $\lambda$ and the observed inverse
document frequency \((IDF)\) as follows:

\[
\lambda = \frac{cf}{N}, \quad IDF = \log_2 \frac{N}{df}
\]

\[
\hat{\alpha} = \frac{\lambda}{\beta} = \frac{cf}{N} \times \frac{df}{df - df(2)}
\]

\[
\hat{\beta} = \lambda \times 2^{IDF} - 1 = \frac{cf - df}{df} = \frac{cdf(2)}{df}
\]

The parameter \(\alpha\) captures the absolute frequency of the term. Two terms with the same \(\beta\) have identical ratios of collection frequency to document frequency, but different values for \(\alpha\) if their collection frequencies are different. The parameter \(\beta\) is the number of “extra terms” per document in which the term occurs (compared to the case where a term has only one occurrence per document). The decay factor \(\hat{\beta}/(\hat{\beta} + 1) = \frac{cdf(2)}{cf}\) (extra terms per term occurrence) determines the ratio \(P_{katz}(k+1)/P_{katz}(k)\), which is the quantity used to measure conditional probability, \(P_w(k+1|k)\).

### 4.2 Katz’s three-parameter model

The conventional Katz K mixture model is called a two-parameter model in Katz’s paper. He pointed out that the two-parameter model does not have enough parameters to treat \(P_w(2|1)\) differently from the rest of the conditional probabilities, \(P_w(k+1|k), k \geq 2\). Accordingly, the three-parameter model, where the difference between \(P_w(2|1)\) and the rest of conditional probabilities, \(P_w(k+1|k)\) for all \(k \geq 2\) are considered, is also proposed in his paper. The model descriptions are shown as follows (Katz 1996):

\[
P_{Katz3}(tf(w, X) = k) = (1 - \alpha)\delta_{k,0} + \alpha \times (1 - \gamma) \times \delta_{k,1} + \frac{\alpha \times \gamma}{\beta + 1}(\frac{\beta}{\beta + 1})^{k-2} \times (1 - \delta_{k,0} - \delta_{k,1})
\]

where:

\[
\delta_{i,j} = \begin{cases} 
1 & \text{iff } i = j \\
0 & \text{otherwise}
\end{cases}
\]

The parameter \(\alpha, \gamma\) and \(\beta\) are estimated by:

\[
\hat{\alpha} = \frac{df(1)}{N}
\]

\[
\hat{\gamma} = \frac{df(2)}{df(1)}
\]

\[
\hat{\beta} = \frac{df(1) - df(2)}{df(2)} = \frac{cdf(3)}{df(2)}
\]

We can denote the two Katz models as \(P_{katz}(k; \alpha, \beta)\) and \(P_{katz}(k; \alpha, \gamma, \beta)\), respectively. According to the experimental reports in Katz’s paper, the three-parameter model, called the G-model by Katz, works much better than the well-known conventional Katz K mixture. In the G-model, the frequencies of zero, single and multiple occurrences are treated separately and consequently, the estimations for the \(ddf(0)\) and \(ddf(1)\), computed by \(ddf(0) = N \times P_{katz}(0; \alpha, \gamma, \beta)\) and \(ddf(1) = N \times P_{katz}(1; \alpha, \gamma, \beta)\), respectively, are always
fitted exactly by such construction. The decay factor, $\frac{\beta}{(\beta + 1)}$, for representing the remaining conditional probabilities is equal to $\frac{\text{cdf}(\beta)}{\text{cdf}(2)}$. To illustrate the problems of the K mixture for modeling term distribution, let us first look at some examples. The number of documents with $k$ (from 1 to 9) occurrences in the test corpus for both six function words and six content-bearing words are presented in Table 1. In the cases of function words, we found that the two K mixture models are both able to model a reasonable frequency distribution, and there is no great difference between their estimations. However, for content-bearing words, the situation is less favorable. Two Katz models behave differently. The conventional model is far worse than the three-parameter model. The estimations by the three-parameter model are quite reasonable for the frequencies of the tail in the distribution, but comparatively serious misjudgments of frequencies occur in the beginning of the distribution except for the points 0 and 1. To our understanding, the kernel part of the K mixture model lies in the method to decide the decay factor. The more precisely the conditional probabilities, $P_w(k + 1|k), k \geq 1$, are reflected by the decay factor, the better the predictions of $ddf(k)$ will be. But what is the characteristic of the conditional probabilities, $P_w(k + 1|k)$, in the case of a different $k$? What are the differences among some possible decay factors, and how do they influence the prediction results? Answers to these questions will help find improvements of the K mixture model and make it more attractive for content-bearing words.

5 Observations

5.1 Observation on the conditional probabilities

The basic assumption of the K mixture model is that the conditional probability of a term is independent of the number $k$ and approximates them by some constant. How is the constant determined empirically based on the assumption of K mixture? Although there are a number of equivalent ways to explain this, we have chosen the following method to better explain why the process goes beyond the K mixture model.

Given the assumption:
\[ P_w(k + 1|k) = C; \]

Empirical estimation of the conditional probability given $k$:
\[ \hat{P}_w(k + 1|k) = \frac{P(tf(w,X) \geq k+1)}{P(tf(w,X) \geq k)} = \frac{df(2)}{df(1)} \]
\[ \hat{P}_w(3|2) = \frac{P(tf(w,X) \geq 3)}{P(tf(w,X) \geq 2)} = \frac{df(3)}{df(2)} \]
\[ \ldots \]
\[ \hat{P}_w(k + 1|k) = \frac{P(tf(w,X) \geq k+1)}{P(tf(w,X) \geq k)} = \frac{df(k+1)}{df(k)} \]
For the conventional K mixture model:
\[
\frac{df(2)}{df(1)} = \frac{df(3)}{df(2)} = \cdots = \frac{df(k+1)}{df(k)} = \frac{\beta}{\beta + 1} = C
\]
\[
\Rightarrow C = \frac{df(2) + df(3) + \cdots + df(k+1)}{df(1) + df(2) + \cdots + df(k)} = \frac{c - df(1)}{c} = \frac{cdf(2)}{cdf(1)}
\]

For the three parameter model:
\[
\frac{df(3)}{df(2)} = \cdots = \frac{df(k+1)}{df(k)} = \frac{\beta}{\beta + 1} = C'
\]
\[
\Rightarrow C' = \frac{df(3) + \cdots + df(k+1)}{df(2) + \cdots + df(k)} = \frac{c - df(1) - df(2)}{c - df(1)} = \frac{cdf(3)}{cdf(2)}
\]

It is apparent that the derived value \(cdf(2)/cdf(1)\) in the conventional K mixture model does not differentiate the case of \(P_w(2|1)\) from the rest of the conditional probabilities \(P_w(k + 1|k), k \geq 2\) while the Katz's three-parameter model does. As pointed out by Katz, in practice, the conditional probability \(P_w(2|1)\) is an entity that is different from the other conditional probabilities, \(P_w(k + 1|k), k \geq 2\), since they are related to different properties. \(P_w(2|1)\) is about how likely a word is to be used again once it first occurs, while the latter are about the probabilities of consecutive repeats provided that the word had already been used repeatedly. Meanwhile, intuitively speaking, a word having occurred, does not necessarily tend to repeat, whereas, an often-repeated word will probably be repeated again. The comparison of conditional probabilities under different \(k\) values for some content-bearing words and function words with notable characteristic are plotted in Figure 1. The distinct phenomenona are: for the content-bearing words, \(P_w(2|1)\) are generally less than \(P_w(k + 1|k)\) and when \(k \geq 2\), \(P_w(k + 1|k)\) is tend to stabilize with the increment of \(k\), while for the function words, there are no great differences in \(P_w(2|1)\) from the rest of the conditional probabilities. This could be explained by the linguistic background. The cause of repetitions of function words is the requirement of forming an integrated language structure for a document. Since the structure of a language is the common rule no matter how many times a word has already been repeated in all kinds of documents, the repetitive features of such types of words will likely be captured by the frequency statistics from large-scale text collection, and where the variance degree of the conditional probabilities is somewhat small. However, the cause of repetitions of content-bearing words is probably determined by the requirements of the author himself to ensure that the meaning of the document is clear. Particularly, the conditional probability reflecting the repetition characteristics will become larger and more stable with an increment in the number of instances of a content-bearing word in a document due to its tendency to be a topic-related word. That is to say, the single occurrence of a word may not cover the topic of the document while a word with multiple occurrences will probably be a topic-related word. Accordingly, it is reasonable for the Katz three-parameter model to treat the \(P_w(2|1)\) and \(P_w(k + 1|k), k \geq 2\) differently. However, accidental repeats of non-topically used content words are possible, par-
Fig. 1 Comparison of the conditional probability $Pr(k + 1|k)$ for $1 \leq k \leq 9$ using sample words with fat-tail properties, including eight content-bearing words and eight function words. There are only slight differences in the conditional probabilities for function words, while, for content-bearing words, there are clear changes between $P(3|2)$ and $P(2|1)$. 
ticularly for the frequently-used content words or in a long document. This is overlooked in
the Katz three-parameter model since the conditional probabilities, $P_w(k + 1|k), k \geq 2$, are
treated equally without discrimination. To further investigate the conditional probabilities,
experiments were conducted on 289 content-bearing words. The mean of $P_w(k + 1|k), k \geq 1$
and other advanced descriptive statistics including $P_{25}, P_{75}$ and Median are all plotted in
Figure 2. In addition to the phenomenon of $P_w(2|1) < P_w(k + 1|k), k \geq 2$, we note that
$P_w(3|2)$ is also comparatively lower than the other conditional probabilities. This charac-
teristic can be explained by the accidental repeats of content-bearing words. Considering the
comparison between $P_w(3|2)$ and $P_w(4|3)$, they are empirically estimated by $df(3)/df(2)$ and
$df(4)/df(3)$, respectively. Without loss of generality, we assume that the accidental repeats
are all accounted for in $df(2)$, $df(3)$ and $df(4)$. From the linguistic intuitions, it is reasonable
to say that the accidental components in $df(2)$ are the largest among them, and the acciden-
tal components of $df(3)$ are larger than $df(4)$. Also, the difference in accidental components
between $df(2)$ and $df(3)$ is bigger than that between $df(3)$ and $df(4)$. Accordingly, we see
that $P_w(3|2)$ is less than $P_w(4|3)$. Moreover, since the occurrence possibility of the accidental
repeats phenomenon will become smaller with the increment of instances of a content-bearing
word in a document, the accidental components in the count for $df(k)$ tends to be smaller, and
topic-repetitions will play a dominant role in determining the value of $P_w(k + 1|k)$. Therefore,
the plots in the figure shows that the conditional probabilities rise initially and then tend to
become stable gradually.

5.2 Observations on decay factors

After learning about the characteristics of the conditional probabilities, let us now study
the relationship of the decay factors with the conditional probabilities, and how the decay
factors are used in the Katz models to reflect conditional probabilities. Now, going back to
the derivations in Section 5.1, the equations for calculating the constant $C$ can be expressed
as follows:

$$C = \frac{\beta}{\beta + 1} = \frac{cdf(2)}{cdf(1)} = \frac{df(2)}{df(1)} + \sum_{k \geq 3} \frac{df(k)}{df(1) + \sum_{k \geq 2} df(k)}$$

Then, we decompose the conventional ratio into two separated ratio factors:

$$C' = \frac{df(2)}{df(1)} \quad \text{and} \quad C'' = \frac{\sum_{k \geq 3} df(k)}{\sum_{k \geq 2} df(k)} = \frac{cdf(3)}{cdf(2)}$$

$C'$ is the conditional probability corresponding to $P_w(2|1)$. As for $C''$, as used in the Katz
three-parameter model, it is the average value of the conditional probabilities $P_w(k + 1|k)$ for
all $k \geq 2$. The initial definition of this value given by Katz is that this average represents a

1 $P_{25}$ and $P_{75}$ means 25th percentiles and 75th percentiles, respectively

142
Xu and Umemura

Improvements of Katz K mixture model

Fig. 2 The comparison of conditional probability $Pr(k+1|k)$ for $1 \leq k \leq 9$. There are 289 content-bearing words with fat-tail properties used in the experiments.

ratio of the expected number of repeats to the expected number of instances that could have been followed by the repeats. We are expecting that there are different properties among the three possible measures $C$, $C'$ and $C''$ for the conditional probabilities of a given word. Therefore, the question arises: What is the practical comparison among $cdf(2)/cdf(1)$, $df(2)/df(1)$ and $cdf(3)/cdf(2)$?

Our investigations were also conducted on the 289 content-bearing words used in Section 5.1. The mean and other advanced descriptive statistics of $cdf(2)/cdf(1)$ and $cdf(3)/cdf(2)$ for these words are appended in Figure 2. The comparisons among the three possible decay factors are: the value of $cdf(3)/cdf(2)$ is larger than $df(2)/df(1)$, and the values of $cdf(2)/cdf(1)$ and $df(3)/df(2)$ are located in between. Meanwhile, plots of the decay factors and the conditional probabilities indicate that: the $P_w(2|1)$ is almost always overestimated,
and the $P_w(k+1|k, k \geq 2)$ are on average underestimated by the decay factor $C$ used in the conventional model. In the three parameter Katz model, although $P_w(2|1)$ is treated individually, $P_w(3|2)$ is overestimated and the remaining conditional probabilities are underestimated by the decay factor $C''$. If the decay factor $df(2)/df(1)$ is selected, $P_w(2|1)$ can be exactly estimated, and all the remaining conditional probabilities will be greatly underestimated. These relationships help to understand the characteristics of the estimation results reported in Section 4. Further comparisons of the decay factors word-by-word for 100 words are also presented in Figure 3. A consistent phenomenon is observed. Among these observations, we are particularly interested in the conditional probability $P_w(3|2)$ and its relationship with those decay factors. The Katz three-parameter model was proposed based on the three-way classification (unrelated, non-topical, topical) of documents and the states of the content-bearing words with occurrences information are only non-topical or topical and are treated by $P_w(2|1)$ and $P_w(k+1|k), k \geq 2$ separately. However, there appears to be a transitive state between the non-topical and topical states. Our intuition is that for a word with a transitive state, its repetitions will be influenced by both non-topical repeats (accidental repeats) and topical repeats. Accordingly, the accidental repeats of such types of words should be considered. Corresponding to practical instances, the conditional probability, $P_w(3|2)$, is something like a transitive state (where the word is), while $P_w(2|1)$ is the non-topical state and $P_w(k+1|k), k \geq 2$ is the topical state.

5.3 Influences of the decay factors on estimation results

In this section, some practical experimental results are reported using the decay factor $C'$ and $C''$, respectively. Instead of treating the non-topical and topical documents differently for a given word, with a slight difference from the Katz three-parameter model, the common decay factors for both non-topical and topical documents are used rather than treating them differently. The method is as follows:

\[
P_{var}(tf(w,X) = k) = (1 - \alpha)\delta_{k,0} + \alpha \gamma \left( \frac{\beta}{\beta+1} \right)^{k-2} \times (1 - \delta_{k,0})
\]

where: $\delta_{k,0} = \begin{cases} 1 & \text{iff } k = 0 \\ 0 & \text{otherwise} \end{cases}$

\[
\hat{\alpha} = \frac{df(1)}{N} \quad \hat{\gamma} = \frac{df(2)}{df(1)}
\]

Variation 1 (Adap): decay factor $\frac{\beta}{\beta+1}$ are estimated by:

\[
\hat{\beta} = \frac{df(3)}{df(1)} \quad \hat{\gamma} = \frac{df(2)}{df(1)}
\]

Variation 2 (B2) : decay factor corresponding to the G-model:

\[
\hat{\beta} = \frac{cf-df(\geq 1)-df(\geq 2)}{cf-df(\geq 1)} = \frac{cdf(3)}{cdf(2)} \quad \hat{\gamma} = \frac{cdf(3)}{df(2)}
\]
Fig. 3 Comparison among $\frac{\text{cdf}(3)}{\text{cdf}(2)}$, $\frac{\text{cdf}(2)}{\text{cdf}(1)}$, $\frac{\text{df}(2)}{\text{df}(1)}$ and $\frac{\text{df}(3)}{\text{df}(2)}$ of the 100 content-bearing words

Since the experimental data is news data selected, to help us find a somewhat consistent phenomenon from the results, the comparison of K mixture variations with the actual values for five content-bearing words is performed on the subset corpus with different time interval, respectively, and results are shown in the Table 2. Meanwhile, the term distribution plots ($k$ from 2 to 9) of four content-bearing words, “congress”, “bush”, “investment” and “japanese” are illustrated as well. The phenomenons what we could find are:

1. In the case of $k = 1$, the variation using the decay factor $\frac{df(2)}{df(1)}$ could exactly estimate the value $ddf(1)$.
2. In the case of $k = 2$, a biased estimation occurred in both the variation “Adap” and the variation “B2”. In most of the cases, the value of $ddf(2)$ will be overestimated by the variation “Adap” while underestimated by the variation “B2”. However, the estimation by the variation “B2” is comparatively close to the actual value.
3. From $k = 4$, the estimation by the variation “B2” approaches the actual value closer than that by variation “Adap”.
4. The decay degree of the variation “Adap” is greater than the variation “B2”.

Fig. 4 Comparison experiments using different decay factor for the words “bush”, “congress”, “investment” and “japanese”. Since the statistic for the “$k = 1$” is fitted exactly in the experiments using $df(2)/df(1)$, the range of the comparison is from 2 to 9. $df(2)/df(1)$ systematically overestimates when $k$ is small, and underestimates when $df(k)$ is large. In contrast, $cdf(2)/cdf(1)$ has the reverse problem. A hybrid of the two should perform better than either by itself.

From this practical comparison, we can conclude that the influences of the decay factor $df(2)/df(1)$ on estimating $ddf(k)$ are negligible for a large $k$ while worthwhile considering for small $k$. It is reasonable for us to expect that the deficiency of the variation “B2” on estimating the frequencies of a small $k$ can be made up by cooperating with the variation “Adap”.

146
<table>
<thead>
<tr>
<th>Corpus</th>
<th>Word</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2 The comparison experiments between the Katz K mixture variations and the actual value for 5 content-bearing words.
6 Our approach

6.1 Model description

The above variations of the Katz K mixture model could be denoted as $P_{Adap}(k; \alpha, \gamma, C')$ and $P_{B2}(k; \alpha, \gamma, C'')$, respectively. Reviewing the comparison of the decay factors and considering the estimation results, it appears that the decay factor, as one deterministic component, is not just a constant, but is in fact a combination of two possible decay factors ($C'$ and $C''$) which is interpolated by a function depending on the number of instances of a word in a document. With this approach, it is possible to achieve somewhat more precise frequency estimation of the transitive state. Accordingly, our proposed model is:

$$P_{our}(tf(w, X) = k) = (1 - \alpha)\delta_{k,0} + \alpha \gamma \left( \frac{\beta_k}{\beta_{k+1}} \right)^{k-\gamma} \times (1 - \delta_{k,0})$$

where:

$$\delta_{k,0} = \begin{cases} 1 & \text{iff } k = 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\alpha = \frac{df(1)}{N}$$

$$\gamma = \frac{df(2)}{df(1)}$$

Decay factor $\frac{\beta_k}{\beta_{k+1}}$ are estimated by:

$$\frac{\beta_k}{\beta_{k+1}} = f(k) \times C' + (1 - f(k)) \times C''$$

$$= f(k) \times \frac{df(2)}{df(1)} + (1 - f(k)) \times \frac{cdf(2)}{cdf(1)}$$

Since factor $C''$ will play a dominant role with the increment of $k$, we choose:

$$f(k) = a \times k^{-b}$$

In our proposed model, in addition to using the three-way classification of documents for counting rate of words, the accidental repeats of non-topically-used content words are also taken into consideration. Experimental results shown above indicate that there is a transitive state between the non-topical and topical state and word repetitions will be influenced by both non-topical repeats and topical repeats and there is possibility to get better estimations using interpolation of two possible decay factors. Therefore the function $f(k)$ is introduced to make the model more linguistic favoritism. It might be possible to form the theoretical models for interpolation factor $f(k)$, and it will be our future work.

In the function $f(k)$, two parameters, $a$ and $b$ need to be ascertained. According to the experimental results of K mixture variations, we know that in the case of $k = 1$, the decay factor will be decided by $df(2)/df(1)$ itself. To make an exact estimation for $ddf(1)$, parameter $a$ should be fixed at 1. Accordingly, only parameter $b$ is required to be processed.

On the basis of admitting the rationality of the K mixture on modeling term distribution, the introduced parameter $b$ is assumed to be chosen so that the number of terms whose frequency distributions fit for the K mixture model should become maximum. For any individual
Table 3 Experiments for 245 content-bearing words using different parameter b based on the three subset corpus (94-95, 96-97, 01-02) with the range of document length from 300 to 700. The number of documents of each subset corpus for 94-95, 96-97 and 01-02 are 78673, 96494 and 45899, respectively. The value in the table is the percentage of terms which pass the χ²-test.

terms, the χ² statistics are chosen for goodness of fit. A χ² statistic, s, for a given word has to be computed for each tested frequency distribution:

\[ s = \sum_{i=1}^{k} \frac{(\text{ddf}(i) - \hat{\text{ddf}}(i))^2}{\hat{\text{ddf}}(i)} \]

where:

\[ \begin{align*}
\text{ddf}(i): & \text{ actual frequencies} \\
\hat{\text{ddf}}(i): & \text{ estimated frequencies; } \hat{\text{ddf}}(i) = N \times P_{\text{our}}(i; \alpha, \gamma, \beta_i(b))
\end{align*} \]

For test reliability, the cells for which the estimated frequencies \( \hat{\text{ddf}}(i) \) are smaller than five have to be combined with adjacent cells so that each combined cell’s estimate is no less than five. In our experiments, because the K mixture is a monotonous function diminishing successively, we can safely combine the cell having the estimated frequency less than five with the remaining cells. The estimation is computed by the identities \( \sum_{r \geq k} C^r = C^k/(1 - C) \).
Meanwhile, the degrees of freedom (d.f.) are adjusted correspondingly. Since the Katz three-parameter model and our proposed K mixture both fit exactly for ddf(0) and ddf(1), the truncated frequency distribution (frequency of the cells at the interval from 2 to 8) is tested. The critical values of the $\chi^2$ test are selected according to a suitable d.f. at the significant level 0.005. If the value of $s$ for a given term is less than the critical value, it will be counted as a term whose frequency distribution could be generated by the hypothesized distribution.

This paper proposes two ways for constructing the models. One is to set up the parameter $b$ as a constant value which lies in the range from 2.0 to 3.0. Since the experimental results in Table 3 show that our approach using a different parameter $b$ all performed better than that of the Katz three-parameter model, an acceptable solution would be that we select a possible value for $b$ in the range from 2.0 to 3.0. From this point of view, the proposed method can be regarded as a kind of three-parameter model. Secondly, to gain much greater improvements, parameter $b$ can be adjusted by a sampling of words. In this paper, the 100 content-bearing words are used as training sets for tuning the parameter $b$. The parameter tuning process is shown in a fragment of the AWK code. After the system running, the most suitable range of parameter $b$ is from 2.5 to 2.8. In the model comparison experiments, the parameter $b$ is fixed at 2.6.

### 6.2 Model comparison

In this section, the comparison experiments among the conventional Katz K mixture, denoted as K’1, Katz three-parameter model denoted as K’2, and our proposed K mixture are conducted.

<table>
<thead>
<tr>
<th>sample words</th>
<th>Document length</th>
<th>#doc</th>
<th>$\chi^2$-test ($k = 2$)</th>
<th>$\chi^2$-test ($k = 2 \sim 8$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>K’1</td>
<td>K’2</td>
</tr>
<tr>
<td>agreement, elections</td>
<td>1-6950(all)</td>
<td>656269</td>
<td>1.6</td>
<td>31.5</td>
</tr>
<tr>
<td>defense, ceasefire</td>
<td>350-400</td>
<td>47767</td>
<td>5.5</td>
<td>83.5</td>
</tr>
<tr>
<td>fighters, terrorism</td>
<td>400-450</td>
<td>32474</td>
<td>10.2</td>
<td>89.0</td>
</tr>
<tr>
<td>germany, italy</td>
<td>600-700</td>
<td>26465</td>
<td>8.7</td>
<td>67.7</td>
</tr>
<tr>
<td>jerusalem, mslm</td>
<td>300-500</td>
<td>153132</td>
<td>1.6</td>
<td>58.3</td>
</tr>
<tr>
<td>parliament, sanctions</td>
<td>500-700</td>
<td>66183</td>
<td>3.1</td>
<td>41.7</td>
</tr>
<tr>
<td>white house, hospital</td>
<td>700-1000</td>
<td>9515</td>
<td>51.2</td>
<td>78.0</td>
</tr>
</tbody>
</table>

**Table 4** Experiments for 127 content-bearing words based on the subset corpus with different range of document length. The value in the table is the percentage of terms which pass the $\chi^2$-test.
Xu and Umemura

Improvements of Katz K mixture model

Function pow(x,y) { z=exp(y*log(x)); return z; }
function x2-critical(word,powv){ # to get critical value for word according the correct degree of freedom.
    freedom_deg=0;
    for(m=2;m<9;m++)
    MIXC=1/pow(m,powv) * C1[word] + (pow(m,powv)-1)/pow(m,powv) * C2[word];
    $\beta = 1/(1-MIXC)-1$;
    edf=N* $\alpha$ [word] * C1[word]/ ( $\beta +1$) * pow(MIXC,m-2);
    if( edf < 5 ) { freedom_deg++; break; }
    freedom_deg++;
}
return crival(freedom_deg);
}
function X2-func(word,powv) { # to get the $\chi^2$-test statistic for each word.
    val=0;
    for(m=2;m<9;m++)
    MIXC=1/pow(m,powv) * C1[word] + (pow(m,powv)-1)/pow(m,powv) * C2[word];
    $\beta = 1/(1-MIXC)-1$;
    edf=N* $\alpha$ [word] * C1[word]/ ( $\beta +1$) * pow(MIXC,m-2);
    if( edf<5 )
        edf = N * ( $\alpha$ [word] * C1[word]/ ( $\beta +1$) * (pow(MIXC,m-2)/(1-MIXC)) ;
        adf = cadf[m,word];
        val+= (edf-cadf) * (edf-cadf)/edf;
        break;
    }
    val+= (edf-adf[m,word]) * (edf-adf[m,word])/edf ;
}
return val;

BEGIN{ # standard critical values of the $\chi^2$-test with d.f. from 1 to 8 at the significant level 0.005
{ ...... # get the empirical statistics of words
    cadf[m,word]=dfn; addf[m,word]=dfn-dfn+1;
    C1[word]=df/df; $\alpha$ [word]=df/N ; C2[word]=cdf/cdf2 ... ...
}
END{
    passmax=1; powv=0.1;
    while (powv<5){
        passcnt=0;
        for(word in train-sets){
            x2-stat=X2-func(word,powv);
            if ( X2-stat < x2-critical(word,powv) ) passcnt++;
        }
    if( passmax < passcnt ) { passmax=passcnt; lower_pow = powv; }
    else if( passmax = = passcnt ) { upper_pow = powv; }
    powv=powv+0.1;
    }
    printf("%6.2f %6.2f %6.2f
",passmax,lower_pow,upper_pow);
}

Fig. 5 AWK code fragment for parameter tuning; Maximum value of parameter $b$, denoted as $powv$ in the program, is set to 5, the initial value and increasing step are all 0.1.
As we mentioned in Section 5.1, the accidental repeats phenomenon will be dramatic in two cases, one where the content word is frequently used and the other where the word occurs in a long document. To investigate the effectiveness of our proposed approach in these two cases, two kinds of experiments were performed. The first experiments for 127 content-bearing words based on the whole and a subset of the corpus with different ranges of document length are conducted, and the results presented in Table 4. The value presented in the column $\chi^2 - test$ is the number of terms passing the statistical test. We note that the range of document length in the corpus does have distinct influences on model effectiveness, and the longer the document is, the worse the estimation results will be. Compared with the Katz three-parameter model, our approach will show dramatic improvements in estimation $ddf(2)$ and slight improvements in the estimation of the tail frequencies; The degree of improvements is clear on a test corpus with a wider range of document length. In the second experiment, 583 content-bearing words, including nouns, biographical names and geographical names, with fat-tail properties are involved, and the corpus with a range of documents from 300 to 500 is used. They are grouped by the $df$ value and comparisons are performed for each group, respectively. Results are presented in Table 5. In the case of the bin with a small $df$ value, our proposed approach and the Katz three-parameter model have the same accuracy but with the increment of the $df$ value of the bin, our proposed approach shows advantages, especially for the estimations of $ddf(2)$. From the above experimental results, it is evident that our proposal can help the $K$ mixture model provide more precise estimations, especially in the case of small $k$. Evidence of the effectiveness of our approach can also be found in Figure 6, where, in addition to the statistical mean, comparisons using the advanced statistics including the median and the 25th and 75th percentiles for 127 content-bearing words are all presented.

<table>
<thead>
<tr>
<th>df-bin</th>
<th>#word</th>
<th>$k = 2$</th>
<th>$\chi^2 - test$</th>
<th>$k = 2 \sim 8$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>K'1</td>
<td>K'2</td>
<td>our</td>
</tr>
<tr>
<td>50 ~ 300</td>
<td>71</td>
<td>60.6</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>300 ~ 600</td>
<td>101</td>
<td>47.5</td>
<td>95.0</td>
<td>99.0</td>
</tr>
<tr>
<td>600 ~ 900</td>
<td>68</td>
<td>25.0</td>
<td>88.0</td>
<td>97.0</td>
</tr>
<tr>
<td>900 ~ 1200</td>
<td>54</td>
<td>18.5</td>
<td>83.3</td>
<td>94.4</td>
</tr>
<tr>
<td>1200 ~ 3000</td>
<td>137</td>
<td>16.1</td>
<td>85.4</td>
<td>99.2</td>
</tr>
<tr>
<td>3000 ~</td>
<td>153</td>
<td>9.2</td>
<td>77.8</td>
<td>96.8</td>
</tr>
</tbody>
</table>

Table 5 583 content-bearing words are grouped according to their $df$ value and comparison experiments are performed for each group. The value in the table is the percentage of terms which pass the $\chi^2 - test$. 

152
Fig. 6  Model comparison using descriptive statistics "Mean", "P25", "P75" and "Median" of 127 sample words.

7 Conclusion

In this paper, an in-depth exploration on the well-known Katz K mixture model is deeply conducted. The interesting phenomena found in the experiments about the characteristics of conditional probabilities of repetition and the exploration of the decay factors are reported in detail. Based on the findings in the experiments, it is possible to use both ends of the distribution to fit models. That is, not only can document frequencies be used when instances of a word are few, but tail probabilities can also be used. Both document frequencies for small instances of a word and the tail probabilities for large instances are rather easy to estimate empirically. Therefore, we propose an effective approach for improving the K mixture model, where the decay factor is the combination of two possible decay factors interpolated by a func-
tion depending on the number of instances of the word in the document. Results show that the proposed model can generate a statistically-significant better estimation of frequencies, especially the frequency estimation for a word with two mentions in a document. In addition, it is shown that the advantages of our approach will become more evident in two cases: modeling the term distribution for the content-bearing word used frequently and modeling the term distribution for the corpus with a wide range of document length.

Acknowledgement

The authors would like to acknowledge Kenneth W. Church for comments and suggestions and to thank Okabe Masayuki and Nakase Kenta for their help on preparing the draft. This work was supported by the 21st Century COE program “Intelligent Human Sensing”, from the Ministry of Education, Culture, Sports, Science and Technology.

Reference


Xu and Umemura

Improvements of Katz K mixture model


Yinghui Xu: Received the B.E. degree from Xi’an Architecture & Technology University, Xi’an, China in 1996 and the M.E. degree from University of Science And Technology Beijing, Beijing, China in 2000. He is now a postgraduate student for doctor degree in the department of Information and Computer Sciences, Toyohashi University of Technology. His research interests include Information Retrieval, Web Mining, Natural Language Processing.

Kyoji Umemura: Received B.E, M.E. and D.E degrees from the University of Tokyo, Tokyo, Japan in 1981, 1983 and 1991, respectively. Currently he is a Professor in the department of Information and Computer Sciences, Toyohashi University of Technology. His research interests include Information Retrieval, Natural Language Processing, Data Mining, Lisp, Symbolic Computation, Compilers and Operating System.

(Received 2005 2, 7 )
(Revised 2005 4, 19 )
(Accepted 2005 5, 23 )