TF-IDuF: A Novel Term-Weighting Scheme for User Modeling based on Users’ Personal Document Collections

Joeran Beel1, 2, Stefan Langer3, Bela Gipp4

1 National Institute of Informatics Tokyo, Digital Content and Media Sciences Division, Japan
2 Trinity College Dublin, School of Computer Science & Statistics, ADAPT Centre, Ireland
3 Otto-von-Guericke University Magdeburg, Department of Computer Science, Germany
4 University of Konstanz, Department of Computer and Information Science, Germany

Abstract
TF-IDF is one of the most popular term-weighting schemes, and is applied by search engines, recommender systems, and user modeling engines. With regard to user modeling and recommender systems, we see two shortcomings of TF-IDF. First, calculating IDF requires access to the document corpus from which recommendations are made. Such access is not always given in a user-modeling or recommender system. Second, TF-IDF ignores information from a user’s personal document collection, which could – so we hypothesize – enhance the user modeling process. In this paper, we introduce TF-IDuF as a term-weighting scheme that does not require access to the general document corpus and that considers information from the users’ personal document collections. We evaluated the effectiveness of TF-IDuF compared to TF-IDF and TF-Only and found that TF-IDF and TF-IDuF perform similarly (click-through rates (CTR) of 5.09% vs. 5.14%), and both are around 25% more effective than TF-Only (CTR of 4.06%) for recommending research papers. Consequently, we conclude that TF-IDuF could be a promising term-weighting scheme, especially when access to the document corpus for recommendations is not possible, and thus classic IDF cannot be computed. It is also notable that TF-IDuF and TF-IDF are not exclusive, so that both metrics may be combined to a more effective term-weighting scheme.

Keywords: term weighting, user modeling, tf-idf, tf-iduf, recommender systems,

1 Introduction
Term-weighting schemes are used by search engines and by user-modeling and recommender systems. Search engines use term-weighting schemes to calculate how well a term describes a document’s content, while user-modeling and recommender systems use term-weighting schemes to calculate how well a term describes a user’s information need. One popular term-weighting schemes is TF-IDF.

TF-IDF was introduced by Jones (1972) and contains two components: term frequency (TF) and inverse document frequency (IDF). TF is the frequency with which a term occurs in a document or user model. The rationale is that the more frequently a term occurs, the more likely this term describes a document’s content or user’s information need. IDF reflects the importance of the term by computing the inverse frequency of documents containing the term within the entire corpus of documents to be searched or recommended. The basic assumption is that a term should be given a higher weight if few other documents also contain that term, because rare terms will likely be more representative of a document’s content or user’s interests.

While TF-IDF was originally developed for classic search, TF-IDF is also one of the most popular term-weighting schemes for user modeling and recommender systems. For instance, TF-IDF is used by 83% of surveyed text-based research-paper recommender systems (Beel, Gipp, Langer, & Breitinger, 2015), and the concept of IDF is applied in other domains of recommender systems, and applied not only to terms but also to entities such as citations (Baral & Li, 2016; Bollacker, Lawrence, & Giles, 1998; Christidis & Mentzas, 2013; Davoodi, Kianmehr, & Alsharchi, 2013; Diaby, Viennet, & Launay, 2013; Lin et al., 2016; Maiga, Hamou-Lhadj, & Larsson, 2014; Philip, Shola, & Ovye, 2014; Ruotsalo et al., 2013; Wang, Abel, Barthès, & Negre, 2014; Yuan, Zheng, Zhang, & Xie, 2013).

1 Other common abbreviations include TF*IDF, TF–IDF, TFIDF, TF×IDF, and TF×IDF
In our research, we focus on the scenario of user modeling and recommender systems. A typical user-modeling and recommendation process utilizing TF-IDF consists of the following steps (Figure 1).

1. User $u$ possesses a document collection $c_u$. This collection might contain, for instance, all documents that the user downloaded, bought, or read.
2. The user-modeling engine identifies those documents from $c_u$ that are relevant for modeling the user’s information need. Relevant documents could be, for instance, documents that the user downloaded or bought in the past $x$ days. The engine selects these documents as a temporary document collection $c_{um}$ to be used for user modeling.
3. The user-modeling engine weights each term that occurs in $c_{um}$ with TF-IDF

$$TF-IDF = tf(t) \times \frac{N_r}{n_r}$$

$t$: Term to weight
$tf(t)$: Frequency of $t$ in the documents of $c_{um}$
$c_r$: A corpus of documents that may be recommended to $u$
$N_r$: Number of documents in $c_r$
$n_r$: Number of documents in $c_r$ that contain $t$

4. The user-modeling engine stores the $z$ highest weighted terms as user model $um$. These terms are meant to represent the user’s information need.
5. The recommender system matches $um$ with the documents in $c_r$ and recommends the most relevant recommendation candidates to $u$.

![Diagram](diagram.png)

**Figure 1.** Document recommendation and user modeling process with TF-IDF and TF-IDF

TF-IDF is commonly more effective than term frequency alone (Manning, Raghavan, & Schütze, 2009), and there has been much research and discussion on TF-IDF, including various extensions and alternatives (Chen, Weinberger, Sha, & others, 2013; Domeniconi, Moro, Pasolini, & Sartori, 2015; Karisani, Rahgozar, & Oroumchian, 2016; Rousseau & Vazirgiannis, 2013; Wu, Luk, Wong, & Kwok, 2008). For instance, Hiemstra (2000) and Robertson (2004) discussed the theoretical foundation underlying TF-IDF and provided a probabilistic justification. Altincay & Erenel (2010) provide a more detailed overview of weighting schemes. With respect to user modeling, we see two limitations of TF-IDF.

1. To calculate IDF, access to the recommendation corpus is needed, which is not always available. For instance, Nascimento, Laender, Silva, & Gonçalves (2011) create user models locally in their literature recommender system and then send the user model as a search query to the ACM Digital Library (the search results are presented as recommendations). In such a scenario, IDF cannot be calculated by the recommender system.

2. TF-IDF calculates term weights based on TF in the documents selected for the user-modeling process and IDF based on the number of documents containing the terms in the recommendation corpus. The documents in a user’s document collection that are not selected in the user modeling process are ignored (in Figure 1 these documents are the grey documents in $c_u$). However, we assume that these remaining documents contain valuable information, as we will explain in detail later.
In this paper, we introduce TF-IDuF, a term-weighting scheme that addresses the two problems, i.e. TF-IDuF can be calculated without access to the recommendation corpus, and it considers the entire document collection of a user\(^2\). Our research goal is to explain the concept of TF-IDuF and to analyze how TF-IDuF performs compared to TF-IDF and term frequency only (TF-Only). It should be noted that we do not suggest to use TF-IDuF as an alternative to TF-IDF, but instead as a complement. In the future, a combination of TF-IDF and TF-IDuF would be possible.

2 TF-IDuF

The term frequency (TF) component in TF-IDuF is the same as in TF-IDF: terms are weighted higher, the more often they occur in the documents selected for building the user model. However, our user-focused inverse document frequency (IDuF) differs from the classic IDF. While the classic IDF is calculated using the document frequencies in the recommendation corpus, IDuF is calculated using the document frequencies in a user’s personal document collection \(c_u\), where terms are weighted more strongly, the fewer documents in a user’s collection contain these terms.

\[
TF-IDuF = tf(t) \cdot \log \frac{N_u}{n_u}
\]

- \(t\): Term to weight
- \(tf(t)\): Frequency of \(t\) in the documents of \(c_{um}\)
- \(c_u\): A user’s collection of documents
- \(N_u\): Number of documents in \(c_u\)
- \(n_u\): Number of documents in \(c_u\) that contain \(t\)

We now illustrate the rationale behind TF-IDuF with two examples.

Example 1 (see left image in Figure 2): The user modeling engine selects documents \(d_1, d_2, \ldots, d_n\) for the user modeling process. \(d_1\) contains term \(t_1\), and \(d_2, \ldots, d_n\) contain term \(t_2\). The overall term frequency for \(t_1\) and \(t_2\) in \(c_{um}\) is the same. Consequently, the density of \(t_1\) in \(d_1\) must be higher than the density of \(t_2\) in each of the documents \(d_2, \ldots, d_n\). In other words, \(t_1\) occurs very frequently in \(d_1\), while \(t_2\) occurs only a few times in each of the documents \(d_2, \ldots, d_n\). We would therefore assume that \(d_1\) covers \(t_1\) in depth, while \(d_2, \ldots, d_n\) cover the topic \(t_2\) only to some extent. We hypothesize that in this scenario, \(t_1\) is more suitable for describing the user’s information need. Hence, \(t_1\) should be weighted more strongly than \(t_2\), which is the case when using TF-IDuF, since only one document in \(c_u\) contains \(t_1\), while many documents contain \(t_2\).

Example 2 (see right image in Figure 2): The user-modeling engine selects a user’s two most recently downloaded documents \(d_1\) and \(d_2\). \(d_1\) contains \(t_1\) in the same frequency as \(d_2\) contains \(t_2\). Based on term frequency alone, both terms would be considered equally suitable for describing the user’s information need. However, the user’s document collection contains a number of additional documents that contain \(t_2\), but these documents were not selected for the user modeling process, e.g. because they were

---

\(^2\) TF-IDuF was first presented in the PhD thesis of Beel (2015). This paper represents the first peer-reviewed publication.
downloaded many months ago. There are no further documents that contain $t_1$ in the user’s document collection. In this scenario, we may assume that $t_1$ describes a new topic that the author was previously not interested in. We hypothesize that in such a scenario, $t_1$ should be weighted more strongly than $t_2$ because:

- Users are likely to favor recommendations for the newer topic $t_1$ rather than for the older topic $t_2$.
- It is easier to generate good recommendations for $t_1$ than for $t_2$ because there are potentially more documents on $t_1$ that the user does not yet know about compared to documents on $t_2$.
- Users have probably received recommendations for $t_2$ in the past, but they have likely not yet received many recommendations for $t_1$. Hence, for $t_2$, the most relevant documents probably have already been recommended to the user.

3 Methodology

We evaluated TF-IDF with an A/B Test in Docear’s research-paper recommender system. Docear is a reference manager that allows users to manage references and PDF files, similar to Mendeley and Zotero (Beel, Gipp, & Mueller, 2009; Beel, Langer, Genzmehr, & Nürnberger, 2013; Beel, Langer, Gipp, & Nürnberger, 2014). One key difference is that Docear’s users manage their data in mind-maps (Beel, Gipp, Langer, & Genzmehr, 2011). Users’ mind-maps contain links to PDFs, as well as the user’s annotations made within those PDFs. To calculate TF-IDF, each mind map of a user was considered as one document.

Docear’s recommender system calculated term weights for user models with: a) TF-IDF, b) TF-IDF and c) TF-only. We compared the effectiveness of the three approaches as measured by user click-through rates (CTR). The rationale of click-through rate is that the term-weighting approach with the highest CTR is the more effective one. For instance, when we report that TF-IDF had a CTR of 5.14%, this means 5.14% of the 42,888 recommendations created using TF-IDF were clicked.

In the recommender-system community, there is a discussion about the appropriateness of different evaluation metrics, and CTR is sometimes criticized. However, in a recent study, we compared CTR with other evaluation metrics such as precision, nDCG, and user ratings, and concluded that CTR is a sensible metric for our scenario (Beel, Breitinger, Langer, Lommatsch, & Gipp, 2016; Beel & Langer, 2015). Docear’s recommender system displayed 228,762 text-based recommendations to 3,483 users between January – September 2014. All reported results are statistically significant (p<0.05), if not stated otherwise. The recommendation corpus contained around 2 million documents in full-text, most of them in English and from various disciplines. For more details on Docear’s recommender system please refer to Beel, Langer, Kapitsaki, Breitinger, & Gipp (2015), Beel (2015), Beel et al. (2014) and Langer & Beel (2014).

4 Results & Interpretation

Click-through rate for TF-IDF was significantly higher than for TF-Only (5.09% vs. 4.06%), i.e. TF-IDF was approximately 25% more effective than TF-Only (Figure 3). This result confirms the previous findings of TF-IDF being more effective than term frequency alone. Although, this result is not surprising, we are, to the best of our knowledge, the first to empirically confirm this result for research-paper recommender systems.

TF-IDF achieved a CTR of 5.14%, meaning it performed equally well as TF-IDF, with its average CTR of 5.09% (the difference is statistically not significant). While this result is already encouraging, we also analyzed the CTRs of TF, TF-IDF, and TF-IDF taking into account the time since a user was registered (Figure 4). The idea was that if a user had been using Docear for a long time, there would be a higher chance for concept drift, and hence TF-IDF should be more effective, compared to short term users. When looking at Figure 4, one can see that CTR slightly decreases over the first couple of months for all three weighting schemes, and then slightly increases again. We have observed this trend before and identified several potential explanations, which we described in Beel, Langer, et al. (2015) and Beel (2015). With regard to the weighting schemes’ effectiveness, Figure 4 shows that, as expected, TF-only consistently performed worse than TF-IDF and TF-IDF.
TF-IDuF was slightly more effective than TF-Only during the first month. Again, this was to be expected because during the first month, concept drift for users is rather unlikely, and users have rather few documents in their collection (of which the majority is used for the user modeling process). Consequently, TF-IDF was the most effective weighting scheme during the first month (7.03%). During months 2 to 5, TF-IDuF outperformed TF-IDF, which could be seen as an indication that after a few months, Docear’s users begin shifting their focus. In the following months, both weighting schemes perform similarly well.

Discussion & Outlook

Overall, we were positively surprised by the results. We expected TF-IDuF to outperform TF-only, but we did not expect it to be equally effective as TF-IDF. In addition, TF-IDuF even appeared to be more effective than TF-IDF after the first month, which should be analyzed in more detail in the future. Considering that TF-IDuF is faster to calculate than TF-IDF and that TF-IDuF can be calculated locally, without access to the global recommendation corpus, we believe that TF-IDuF can be a valuable weighting scheme. TF-IDuF and TF-IDF are not exclusive and could be used in a complementary manner. This means, a term could be weighted based on all three factors TF, IDF, and IDuF. Further research is necessary, to assess the performance of such a combined TF-IDF-IDuF weighting scheme. In this paper, we performed the first evaluation of TF-IDuF using the mind maps of Docear’s users as personal document corpora. Further research is necessary to confirm the promising performance and to find out if TF-IDuF performs equally well on other types of personal document corpora, such as users’ collections of research-papers, websites or news articles.
6 References


Additional Information

Bibliographic Data

Preprint

Related Work
https://www.gipp.com/pub/

Authors
Joeran Beel
Stefan Langer
Bela Gipp

BibTeX
@InProceedings{Beel17,
  Title = {TF-IDuF: A Novel Term-Weighting Scheme for User Modeling based on Users’ Personal Document Collections},
  Author = {Beel, Joeran and Langer, Stefan and Gipp, Bela},
  Booktitle = {Proceedings of the iConference 2017},
  Year = {2017},
  Address = {Wuhan, China},
  Month = {March 22-25},
  Url = {http://ischools.org/the-iconeference/}
}

RefMan (RIS)
ER -
TY - CONF
AU - Beel, Joeran
AU - Langer, Stefan
AU - Gipp, Bela
T1 - TF-IDuF: A Novel Term-Weighting Scheme for User Modeling based on Users' Personal Document Collections
T2 - Proceedings of the iConference 2017
A0 - Wuhan, China
Y1 - 2017/March 22-25
UR - http://ischools.org/the-iconeference/

EndNote
%0 Conference Proceedings
%A Beel, Joeran
%A Langer, Stefan
%A Gipp, Bela
%T TF-IDuF: A Novel Term-Weighting Scheme for User Modeling based on Users’ Personal Document Collections
%B Proceedings of the iConference 2017
%P 2017/March 22-25
%U http://ischools.org/the-iconeference/
%A Wuhan, China