

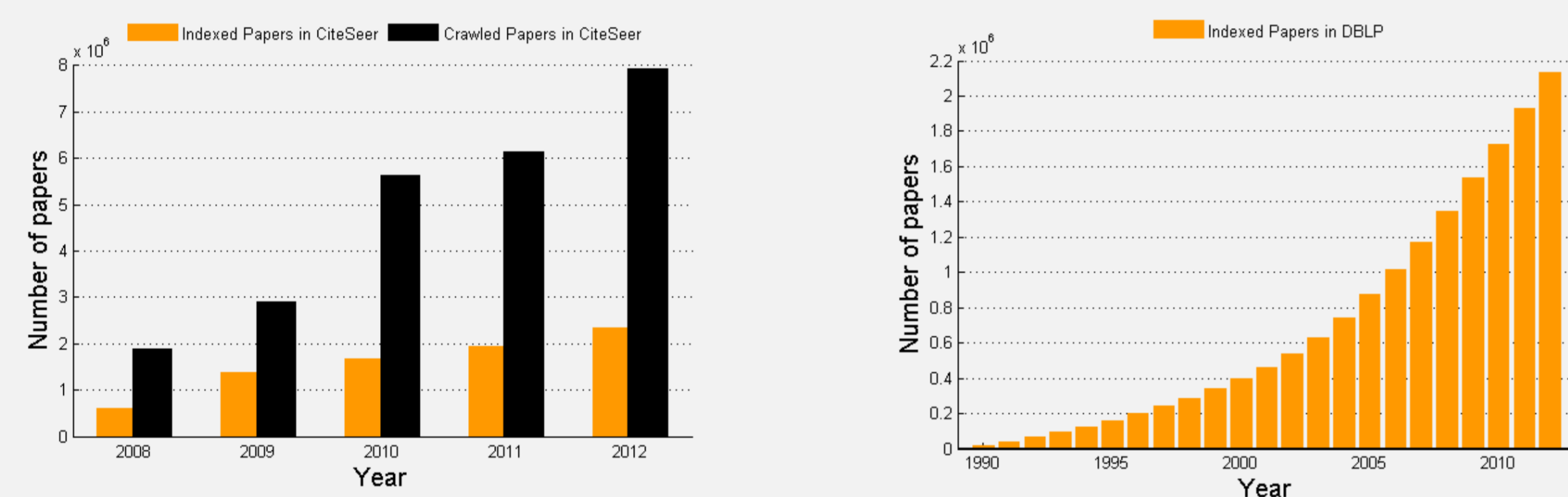
## WHY KEYPHRASE EXTRACTION?

- ▶ The number of scholarly documents on the Web is exponentially increasing every year.
- ▶ Keyphrase extraction is the problem of automatically extracting important phrases or concepts (i.e. the *essence*) of a document.
- ▶ Keyphrases are shown to be rich sources of information for many applications such as document classification, clustering, recommendation, indexing, searching and summarization.
- ▶ During these *big data* times, keyphrases associated with research papers can allow for *efficient processing of more information in less time*.
  - ▶ Also useful in many ML and IR applications such as topic tracking, information filtering, and search.



## STATS

- ▶ Number of Research Papers indexed by two important digital libraries in the fields of computer and information science over the past years.



## FROM DATA TO KNOWLEDGE

- ▶ Scientific research papers typically propose innovative solutions or extend the state-of-the-art algorithms for existing research problems.
- ▶ In addition to a document's textual content, other informative neighborhoods exist that have the potential to improve keyphrase extraction. For example, research papers are highly inter-connected in giant *citation networks*, where papers *cite* or *are cited* by others.
  - ▶ In a citation network, information flows from one paper to another via the citation relation.
  - ▶ These contexts are not arbitrary, but they serve as brief summaries of a cited paper.

**Cited Context**

**Where You Like to Go Next:**  
**Successive Point-of-Interest Recommendation**  
 Chen Cheng, Haiqin Yang, Michael R. Lyu, Irwin King

methods, e.g. **Bayesian Probabilistic Tensor Factorization (BPTF)** [Xing et al., 2010], **factorized personalized Markov chains (FPMC)** [Rendle et al., 2010] and etc., have been proposed and demonstrated themselves as promising methods in

**FPMC**: this method is proposed in [Rendle et al., 2010], which is a strong baseline **model** embedding users' preference and their **personalized Markov Chain** to provide **next-basket item recommendation**.

our **FPMC-LR** borrows the idea of **factoring personalized Markov chain (FPMC)** for solving the task of **next-basket recommendation** [Rendle et al., 2010], we emphasize on users' movement constraint, i.e., moving around a local region, and

**Target Paper**

**Author-annotated keywords:**  
 basket recommendation, markov chain, matrix factorization

Steffen Rendle, Christoph Freudenthaler, Lars Schmidt-Thieme

**ABSTRACT**

Recommender systems are an important component of many websites. Two of the most popular approaches are based on **matrix factorization (MF)** and **Markov chains (MC)**. MF methods learn the general taste of a user by factorizing the **matrix** over observed **user-item** preferences. On the other hand, **MC** methods model **sequential** behavior by learning a **transition graph** over **items** that is used to predict the next action based on the recent actions of a user. In this paper, we present a method bringing both approaches together. Our method is based on **personalized transition graphs** over underlying **Markov chains**. That means for each user an own **transition matrix** is learned – thus in total the method uses a **transition cube**. As the observations for estimating the transitions are usually very limited, our method factorizes the transition cube with a pairwise interaction **model** which is a special case of the Tucker Decomposition. We show that our **factorized personalized MC (FPMC)** model subsumes both a common **Markov chain** and the normal **matrix factorization model**. For learning the **model** parameters, we introduce an adaptation of the **Bayesian Personalized Ranking (BPR)** framework for **sequential basket data**. Empirically, we show that our **FPMC** model outperforms both the common **matrix factorization** and the **unpersonalized MC** model both learned with and without **factorization**.

Markov chains or recommender systems have been studied by several researchers. Zeng et al. [10] describe a **sequential recommender** based on **Markov chains**. They investigate how to extract

Three recent methods for **item recommendation** are based on the **matrix factorization model** that factorizes the **matrix** of **user-item** correlations. Both [Fu et al. [3] and Pan and Scholz [6] optimize the **factorization** on **user-item pairs (u,i)** where observed pairs are treated as positive and

mining methods to discover **sequential patterns** which are used for **generating recommendations**. Shani et al. [9] introduce a **recommender** based on **Markov decision processes (MDP)** and also a **MC** based **recommender**.

Citing contexts

Can we effectively exploit information available in large inter-linked document networks in order to improve the performance of keyphrase extraction?

## CITATION-ENHANCED KEYPHRASE EXTRACTION (CeKE)

We propose a supervised binary classification model called CeKE, built on a combination of novel features that capture information from citation contexts and existing features from previous works.

- ▶ Generating Candidate Phrases
  - ▶ We first apply parts-of-speech filters and retain only the nouns and adjectives
  - ▶ Porter Stemmer is applied on every word
  - ▶ Words that have contiguous positions in the document are concatenated into *n*-grams
  - ▶ Finally, we eliminate phrases that end with an adjective and the unigrams that are adjectives
- ▶ Features

Feature Name	Description
<b>Existing features for keyphrase extraction</b>	
<i>tf-idf</i>	term frequency * inverse document frequency, computed from a target paper; used in KEA
<i>relativePos</i>	the position of first occurrence of a phrase divided by the total number of tokens; used in KEA and Hulth's methods
POS	the part-of-speech tag of the phrase; used in Hulth's methods
<b>Novel features - Citation Network Based</b>	
<i>inCited</i>	if the phrase occurs in cited contexts
<i>inCiting</i>	if the phrase occurs in citing contexts
<i>citation tf-idf</i>	the <i>tf-idf</i> value of the phrase, computed from the aggregated citation contexts
<b>Novel features - Extensions of Existing Features</b>	
<i>first position</i>	the distance of the first occurrence of a phrase from the beginning of a paper
<i>tf-idf-Over</i>	<i>tf-idf</i> larger than a threshold $\theta$
<i>firstPosUnder</i>	the distance of the first occurrence of a phrase from the beginning of a paper is below some value $\beta$

## DATASETS

- ▶ We compiled two datasets consisting of titles and abstracts from two top-tier machine learning conferences:
  - ▶ World Wide Web (WWW)
  - ▶ Knowledge Discovery and Data Mining (KDD)
- ▶ The *author-annotated* keyphrases were treated as the gold standard.
- ▶ The citation contexts' length was set to 50 words around the citation mention.

Dataset	Num. (#) Papers	Average Cited Ctx.	Average Citing Ctx.	Average Keyphrases	#uni-grams	#bi-grams	#tri-grams
WWW	425	15.45	18.78	4.87	680	1036	247
KDD	365	12.69	19.74	4.03	363	853	189

## RESULTS

How does CeKE compare with the existing supervised models that use only information intrinsic to the data?

Method	WWW			KDD		
	Precision	Recall	F1-score	Precision	Recall	F1-score
<b>Citation - Enhanced (CeKE)</b>	<b>0.227</b>	<b>0.386</b>	<b>0.284</b>	<b>0.213</b>	<b>0.413</b>	<b>0.280</b>
Hulth - <i>n</i> -gram with tags	0.165	0.107	0.129	0.206	0.151	0.172
KEA	0.210	0.146	0.168	0.178	0.124	0.145

Figure: Comparison of CeKE with Hulth's *n*-grams with tags and KEA methods. Hulth's features: POS, relative position, document frequency and collection frequency. KEA's features: *tf-idf* and relative position

## RESULTS

How is our citation-enhanced algorithm comparing with recent unsupervised models?

Method	WWW			KDD		
	Precision	Recall	F1-score	Precision	Recall	F1-score
<b>Citation - Enhanced (CeKE)</b>	<b>0.227</b>	<b>0.386</b>	<b>0.284</b>	<b>0.213</b>	<b>0.413</b>	<b>0.280</b>
TF-IDF - Top 5	0.089	0.100	0.094	0.083	0.102	0.092
TF-IDF - Top 10	0.075	0.169	0.104	0.080	0.203	0.115
TextRank - Top 5	0.058	0.071	0.062	0.051	0.065	0.056
TextRank - Top 10	0.062	0.133	0.081	0.053	0.127	0.072
ExpandRank - 1 neigh. - Top 5	0.088	0.109	0.095	0.077	0.103	0.086
ExpandRank - 1 neigh. - Top 10	0.078	0.165	0.101	0.071	0.177	0.098
ExpandRank - 5 neigh. - Top 5	0.093	0.113	0.100	0.080	0.108	0.090
ExpandRank - 5 neigh. - Top 10	0.080	0.172	0.104	0.068	0.172	0.095
ExpandRank - 10 neigh. - Top 5	0.094	0.113	0.100	0.077	0.103	0.086
ExpandRank - 10 neigh. - Top 10	0.076	0.162	0.099	0.065	0.164	0.091

Figure: Comparison of CeKE state-of-the-art unsupervised systems. TextRank: window size is set to 2. ExpandRank: window size is set to 10.

How well does our proposed model perform in the absence of either cited or citing contexts?

Method	WWW			KDD		
	Precision	Recall	F1-score	Precision	Recall	F1-score
<b>CeKE - Both contexts</b>	<b>0.227</b>	<b>0.386</b>	<b>0.284</b>	<b>0.213</b>	<b>0.413</b>	<b>0.280</b>
CeKE - Only cited contexts	0.222	0.286	0.247	0.192	0.300	0.233
CeKE - Only citing contexts	0.203	0.342	0.253	0.195	0.351	0.250

Figure: Results of CeKE using both context and using only cited or citing contexts

## ANECDOTAL EVIDENCE

- ▶ Our classifier was trained on both WWW and KDD datasets and was evaluated on an award winning paper published in the EMNLP Conference.
- ▶ We gathered from the Web 49 cited contexts and 30 citing contexts.
- ▶ The classifier was tuned to return as keyphrases only those that had an extremely high probability (we used a threshold of 0.985).

**Unsupervised Semantic Parsing<sup>0.997</sup>**

We present the first unsupervised approach to the problem of learning a **semantic parser<sup>1.000</sup>**, using **Markov logic<sup>0.991</sup>**. Our **USP system<sup>0.985</sup>** transforms dependency trees into quasi-logical forms, recursively induces lambda forms from these, and clusters them to abstract away syntactic variations of the same meaning. The MAP **semantic parse<sup>1.000</sup>** of a sentence is obtained by recursively assigning its parts to lambda-form clusters and composing them. We evaluate our approach by using it to extract a knowledge base from biomedical abstracts and answer questions. **USP<sup>1.000</sup>** substantially outperforms TextRunner, DIRT and an informed baseline on both precision and recall on this task.

Human annotated labels: *unsupervised semantic parsing, Markov logic, USP system*

Figure: The title and abstract of an award winning paper published in the EMNLP conference by Poon and Domingos (2009). Grey - filtered out words; Black - candidate phrases; Bold red - predicted keyphrases; Numbers - classifier's confidence.

Keyphrase	#cited c.	#citing c.	
<i>semantic parser</i>	29	26	The table on the left shows the term-frequency of every predicted keyphrase within the citation network.
<i>USP</i>	31	10	
<i>Markov logic</i>	15	10	
<i>unsupervised semantic parsing</i>	12	1	
<i>USP system</i>	3	2	

## CONCLUSIONS AND FUTURE DIRECTIONS

- ▶ The proposed citation-enhanced supervised framework performs substantially better compared with state-of-the-art supervised and unsupervised models.
- ▶ Our model can be extended to other types of documents such as webpages, weblogs or e-mails.
- ▶ Using external sources, e.g. Wikipedia, could increase the performance by identifying better candidate phrases.
- ▶ Extensions to other domains, e.g. Biology and Chemistry, and other applications, e.g. document summarization, are of particular interest.