WWW(ACM) International World Wide Web Conferences

相关论文整理

1. [Question Answering over Pattern-Based User Models](http://dl.acm.org/citation.cfm?id=2993331&CFID=715860528&CFTOKEN=30536297)

12 - 15, 2016

链接：

<http://dl.acm.org/citation.cfm?id=2993331&CFID=715860528&CFTOKEN=30536297>

摘要：

In this paper we present an ontology-driven framework for natural language question analysis and answering over user models (e.g. preferences, habits and health problems of individuals) that are formally captured using ontology design patterns. Pattern-based modelling is extremely useful for capturing n-ary relations in a well-defined and axiomatised manner, but it introduces additional challenges in building NL interfaces for accessing the underlying content. This is mainly due to the encapsulation of domain semantics inside conceptual layers of abstraction (e.g. using reification or container classes) that demand flexible, context-aware approaches for query analysis and interpretation. We describe the coupling of a frame-based formalisation of natural language user utterances with a context-aware query interpretation towards question answering over pattern-based RDF knowledge bases. The proposed framework is part of a human-like socially communicative agent that acts as an intermediate between elderly migrants and care personnel, assisting the latter to solicit personal information about care recipients (e.g. medical history, care needs, preferences, routines, habits, etc.).

在本文中，我们提出一个本体驱动的框架，用于自然语言问题分析和回答用户模型（例如偏好，习惯和个人的健康问题），使用本体设计模式正式捕获。基于模式的建模对于以良好定义和公理化的方式捕获n元关系非常有用，但它在构建用于访问基础内容的NL接口时引入了额外的挑战。这主要是由于在抽象的概念层（例如使用引用或容器类）内部需要灵活的上下文感知方法来进行查询分析和解释的域语义的封装。我们描述了基于帧的自然语言用户话语形式化与上下文感知查询解释到基于模式的RDF知识库的问题回答的耦合。拟议的框架是人类社会交际代理的一部分，作为老年移民和护理人员之间的中间人，帮助他们征求关于照顾者的个人信息（例如病史，护理需要，喜好，习惯，习惯等。）。

**关键词**：语言分析，问题解答，用户模型，本体设计模式

**提出的问题：**

提出一个本体驱动的框架，用于自然语言问题分析和回答用户模型（例如偏好，习惯和个人的健康问题）。

**解决方法：**

使用本体设计模式正式捕获。

# **Semi-Automated Information Extraction from Unstructured Threat Advisories**

05 - 07, 2017

链接：

<http://dl.acm.org/citation.cfm?id=3021482&CFID=715860528&CFTOKEN=30536297>

摘要：

One of the fundamental challenges for information officers of most organizations today is the growing number of cyber security threats. This has led to an emerging field of Cyber Threat Intelligence, which is a mechanism to acquire, categorize and prioritize information regarding impending security threats from disparate online sources, enabling organizations to take the necessary steps to avoid compromising client data and protecting their hardware and software resources. Such information is published as formal security advisories which are largely in the form of unstructured or semi structured data. In this work we describe an approach to read large volume of such unstructured data and automatically extract useful nuggets of information like the exploit targets, techniques for the exploitation and recommended prevention guidelines. We use natural language processing techniques and a pattern identification framework to extract these information nuggets. We present some early results and observations.

今天大多数组织的信息官员的一个根本挑战是网络安全威胁的数量不断增加。这导致了一个新兴的网络威胁智能领域，这是一种从不同的在线来源获取，分类和优先处理即将来临的安全威胁的信息的机制，使组织能够采取必要的措施，避免危及客户数据和保护他们的硬件和软件资源。此类信息作为正式安全建议公布，其形式主要是非结构化或半结构化数据。在这项工作中，我们描述了一种读取大量这种非结构化数据的方法，并自动提取有用的信息块，如利用目标，利用技术和建议的预防指南。我们使用自然语言处理技术和模式识别框架来提取这些信息块。我们提出一些早期的结果和观察。

**关键词**：网络威胁情报，自然语言处理，模式检测，半监督学习，网络安全

**提出的问题：**

从不同的在线来源获取，分类和优先处理即将来临的安全威胁的信息的机制，使组织能够采取必要的措施，避免危及客户数据和保护他们的硬件和软件资源。

**解决方法：**

描述了一种读取大量这种非结构化数据的方法，并自动提取有用的信息块，如利用目标，利用技术和建议的预防指南。我们使用自然语言处理技术和模式识别框架来提取这些信息块。

# **Content-based emotion classification in online social networks for Chinese Microblogs**

# February 03, 2017

# 链接：

# <http://dl.acm.org/citation.cfm?id=3014866&CFID=715860528&CFTOKEN=30536297>

# 摘要：

# Recent years, social networks are popular throughout the whole world. In China in particular, more people spend their time on social networks. Sina Weibo, as the most popular microblogs in China, records millions of microblogs from different population. In this paper, we study and understand sentimental feelings of Weibo by methods of mathematical statistics and analysis. Firstly, we propose a novel three-step extract (*NTSE*) algorithm to extract meaningful microblogs. Secondly, we identify the similarity of microblogs sent by specific population. Then, we present the naive Bayes algorithm to classify microblogs into three types: positive, negative or objective. For testing the algorithms, we collect Weibo data from specific population of Sina Weibo to form two datasets: student dataset and profession dataset. Some interesting findings include: i) around 20% microblogs are meaningless; ii) only half of microblogs' contents have expressed emotion; iii) students tend to post microblogs with negative emotion among the emotional trends; ix) six professional persons tend to publish positive microblogs. The results of our experiments show that students in five universities in China are more inclined to express negative feelings in the social networks. On contrary, some professional persons including IT, actors and writers and so on more likely to publish positive microblogs.

# 近年来，社交网络在全世界流行。特别是在中国，更多的人花时间在社交网络上。新浪微博作为中国最受欢迎的微博，记录了来自不同人口的数百万微博。在本文中，我们通过数学统计和分析的方法研究和理解微博的感性情感。首先，我们提出一种新的三步提取物（*NTSE*）算法来提取有意义的微博。其次，我们确定特定人口发送的微博的相似性。然后，我们提出朴素贝叶斯算法，将微博分为三类：正，负或客观。为了测试算法，我们从新浪微博的特定人群收集微博数据以形成两个数据集：学生数据集和职业数据集。一些有趣的发现包括：i）大约20％的微博是无意义的; ii）只有一半的微博内容表达了情绪; iii）学生倾向于在情感趋势中发布负面情感的微博; ix）六个专业人士倾向于发布积极的微博。我们的实验结果表明，中国五所大学的学生更倾向于在社交网络中表达负面情绪。相反，一些专业人士，包括IT，演员和作家等更有可能发布积极的微博。

# 关键词：自然语言处理，文本分析，情感分析，积极和消极，社交网络

# 提出的问题：

# 新浪微博作为中国最受欢迎的微博，记录了来自不同人口的数百万微博。如何研究和理解微博的感性情感。

# 解决方法：

# 在本文中，我们通过数学统计和分析的方法研究和理解微博的感性情感。首先，我们提出一种新的三步提取物（*NTSE*）算法来提取有意义的微博。其次，我们确定特定人口发送的微博的相似性。然后，我们提出朴素贝叶斯算法，将微博分为三类：正，负或客观。

1. aNMM: Ranking Short Answer Texts with Attention-Based Neural Matching Model

October 24 - 28, 2016

连接：

<http://dl.acm.org/citation.cfm?id=2983818&CFID=715860528&CFTOKEN=30536297>

摘要：

As an alternative to question answering methods based on feature engineering, deep learning approaches such as convolutional neural networks (CNNs) and Long Short-Term Memory Models (LSTMs) have recently been proposed for semantic matching of questions and answers. To achieve good results, however, these models have been combined with additional features such as word overlap or BM25 scores. Without this combination, these models perform significantly worse than methods based on linguistic feature engineering. In this paper, we propose an attention based neural matching model for ranking short answer text. We adopt value-shared weighting scheme instead of position-shared weighting scheme for combining different matching signals and incorporate question term importance learning using question attention network. Using the popular benchmark TREC QA data, we show that the relatively simple aNMM model can significantly outperform other neural network models that have been used for the question answering task, and is competitive with models that are combined with additional features. When aNMM is combined with additional features, it outperforms all baselines.

作为基于特征工程的问题回答方法的替代方案，最近已经提出了用于问题和答案的语义匹配的深度学习方法，例如卷积神经网络（CNN）和长短期存储器模型（LSTM）。然而，为了获得好的结果，这些模型已经与诸如单词重叠或BM25得分的附加特征相结合。没有这种组合，这些模型的表现明显劣于基于语言特征工程的方法。在本文中，我们提出了一种基于注意的神经匹配模型，用于对短篇答案文本进行排名。我们采用价值共享加权方案，而不是位置共享加权方案，用于组合不同的匹配信号，并结合使用问题关注网络的问题词重要性学习。使用流行的基准TREC QA数据，我们显示相对简单的aNMM模型可以显着优于已经用于问题回答任务的其他神经网络模型，并且与结合附加特征的模型竞争。当NMM与其他功能组合时，它的性能优于所有基线。

**关键词**：长期学习的重要性，价值共享的权重，问题解答，深度学习

**提出的问题：**

最近已经提出了用于问题和答案的语义匹配的深度学习方法，例如卷积神经网络（CNN）和长短期存储器模型（LSTM）。然而，为了获得好的结果，这些模型已经与诸如单词重叠或BM25得分的附加特征相结合。没有这种组合，这些模型的表现明显劣于基于语言特征工程的方法。

**解决方法：**

在本文中，我们提出了一种基于注意的神经匹配模型，用于对短篇答案文本进行排名。我们采用价值共享加权方案，而不是位置共享加权方案，用于组合不同的匹配信号，并结合使用问题关注网络的问题词重要性学习。

1. [Improving Unsupervised Dependency Parsing with Knowledge from Query Logs](http://dl.acm.org/citation.cfm?id=2903720&CFID=715860528&CFTOKEN=30536297)

June 2016

连接：

http://dl.acm.org/citation.cfm?id=2903720&dl=ACM&coll=DL&CFID=715860528&CFTOKEN=30536297#URLTOKEN#

Keywords: natural annotations, query logs, additional knowledge, Dependency parsing

摘要：

Unsupervised dependency parsing becomes more and more popular in recent years because it does not need expensive annotations, such as treebanks, which are required for supervised and semi-supervised dependency parsing. However, its accuracy is still far below that of supervised dependency parsers, partly due to the fact that their parsing model is insufficient to capture linguistic phenomena underlying texts. The performance for unsupervised dependency parsing can be improved by mining knowledge from the texts and by incorporating it into the model. In this article, syntactic knowledge is acquired from query logs to help estimate better probabilities in dependency models with valence. The proposed method is language independent and obtains an improvement of 4.1% unlabeled accuracy on the Penn Chinese Treebank by utilizing additional dependency relations from the Sogou query logs and Baidu query logs. Morever, experiments show that the proposed model achieves improvements of 8.07% on CoNLL 2007 English using the AOL query logs. We believe query logs are useful sources of syntactic knowledge for many natural language processing (NLP) tasks.

无监督的依赖性解析在近年来变得越来越流行，因为它不需要昂贵的注释，例如监督和半监督依赖性解析所需的树状图。然而，其精度仍然远低于监督依赖解析器，部分是由于它们的解析模型不足以捕获基础文本的语言现象的事实。通过从文本挖掘知识并将其并入模型中，可以提高无监督依赖性解析的性能。在本文中，从查询日志中获取句法知识，以帮助估计价值依赖模型中的更好的概率。提出的方法是语言独立的，并通过利用来自搜狗查询日志和百度查询日志的附加依赖关系，在宾州中国树木银行获得4.1％的未标记精度的改进。更多，实验表明，使用AOL查询日志，提出的模型实现了对CoNLL 2007英语的8.07％的改进。我们认为查询日志是许多自然语言处理（NLP）任务的句法知识的有用来源。

**关键词：**自然注解，查询日志，额外的知识，依赖解析

**提出的问题：**

无监督的依赖性解析在近年来变得越来越流行，然而，其精度仍然远低于监督依赖解析器，部分是由于它们的解析模型不足以捕获基础文本的语言现象的事实。

**解决方法：**

通过从文本挖掘知识并将其并入模型中，可以提高无监督依赖性解析的性能。在本文中，从查询日志中获取句法知识，以帮助估计价值依赖模型中的更好的概率。

# Adaptation of Language Models for SMT Using Neural Networks with Topic Information

# Article No. 19 2016

# 链接：

http://dl.acm.org/citation.cfm?id=2816816

# 摘要：Neural network language models (LMs) are shown to be effective in improving the performance of statistical machine translation (SMT) systems. However, state-of-the-art neural network LMs usually use words before the current position as context and neglect global topic information, which can help machine translation (MT) systems to select better translation candidates from a higher perspective. In this work, we propose improvement of the state-of-the-art feedforward neural language model with topic information. Two main issues need to be tackled when adding topics into neural network LMs for SMT: one is how to incorporate topics to the neural network; the other is how to get target-side topic distribution before translation. We incorporate topics by appending topic distribution to the input layer of a feedforward LM. We adopt a multinomial logistic-regression (MLR) model to predict the target-side topic distribution based on source side information. Moreover, we propose a feedforward neural network model to learn joint representations on the source side for topic prediction. LM experiments demonstrate that the perplexity on validation set can be greatly reduced by the topic-enhanced feedforward LM, and the prediction of target-side topics can be improved dramatically with the MLR model equipped with the joint source representations. A final MT experiment, conducted on a large-scale Chinese--English dataset, shows that our feedforward LM with predicted topics improves the translation performance against a strong baseline.

# 神经网络语言模型（LM）被显示为有效地改善统计机器翻译（SMT）系统的性能。然而，最先进的神经网络LM通常使用在当前位置之前的词作为上下文并忽略全局主题信息，这可以帮助机器翻译（MT）系统从更高的角度选择更好的翻译候选。在这项工作中，我们提出改进的先进的前馈神经语言模型与主题信息。当向SMT的神经网络LM添加主题时需要解决两个主要问题：一个是如何将主题合并到神经网络;另一个是如何在翻译之前获得目标端主题分布。我们通过将主题分布附加到前馈LM的输入层来合并主题。我们采用多项Logistic回归（MLR）模型来预测基于源方信息的目标方主题分布。此外，我们建议一个前馈神经网络模型来学习在源方面的主题预测的联合表示。 LM实验证明，通过主题增强的前馈LM可以大大减少验证集上的困惑，并且可以使用装备有联合源表示的MLR模型来显着改进目标侧主题的预测。最终的MT实验在大规模中文 - 英语数据集上进行，表明我们具有预测主题的前馈LM提高了翻译性能与强基线。

# 关键词：联合表示，前馈神经网络语言模型，主题模型，统计机器翻译，多项Logistic回归

# 提出的问题：

# 最先进的神经网络LM通常使用在当前位置之前的词作为上下文并忽略全局主题信息，这可以帮助机器翻译（MT）系统从更高的角度选择更好的翻译候选。

# 解决方法：

# 在这项工作中，我们提出改进的先进的前馈神经语言模型与主题信息。英语数据集上进行，表明我们具有预测主题的前馈LM提高了翻译性能与强基线。

# **A differential excitation based rotational invariance for convolutional neural networks**

# **链接：**

http://dl.acm.org/citation.cfm?id=3009978&CFID=715860528&CFTOKEN=30536297

# Article No. 70 2016

# 摘要：

# Deep Learning (DL) methods extract complex set of features using architectures containing hierarchical set of layers. The features so learned have high discriminative power and thus represents the input to the network in the most efficient manner. Convolutional Neural Networks (CNN) are one of the deep learning architectures, extracts structural features with little invariance to smaller translational, scaling and other forms of distortions. In this paper, the learning capabilities of CNN's are explored towards providing improvement in rotational invariance to its architecture. We propose a new CNN architecture with an additional layer formed by differential excitation against distance for the improvement of rotational invariance and is called as RICNN. Moreover, we show that the proposed method is giving superior performance towards invariance to rotations against the original CNN architecture (training samples with different orientations are not considered) without disturbing the invariance to smaller translational, scaling and other forms of distortions. Different profiles like training time, testing time and accuracies are evaluated at different percentages of training data for comparing the performance of the proposed configuration with original configuration.

# 深度学习（DL）方法使用包含分层层组的架构来提取复杂的特征集。所学习的特征具有高辨别力，因此以最有效的方式表示对网络的输入。卷积神经网络（CNN）是深度学习架构之一，提取了对较小的平移，缩放和其他形式的失真几乎不变的结构特征。在本文中，CNN的学习能力被探索以提供对其结构的旋转不变性的改进。我们提出一种新的CNN架构与由距离差分激励形成的附加层为改善旋转不变性，称为RICNN。此外，我们显示所提出的方法给予优越的性能对旋转不变性对原始CNN架构（不考虑具有不同方向的训练样本），而不扰乱对较小的平移，缩放和其他形式的失真的不变性。以不同百分比的训练数据评估不同的轮廓，如训练时间，测试时间和精度，以比较所提出的配置与原始配置的性能。

# 关键词：反向传播，旋转不变性，卷积神经网络，深度学习，激发差

# 提出的问题：

# 卷积神经网络（CNN）是深度学习架构之一，提取了对较小的平移，缩放和其他形式的失真几乎不变的结构特征。在本文中，CNN的学习能力被探索以提供对其结构的旋转不变性的改进。

# 解决方法：

# 卷积神经网络（CNN）是深度学习架构之一，提取了对较小的平移，缩放和其他形式的失真几乎不变的结构特征。在本文中，CNN的学习能力被探索以提供对其结构的旋转不变性的改进。

1. Text Classification with Topic-based Word Embedding and Convolutional Neural Networks

# October 02 - 05, 2016

# 链接：

# <http://dl.acm.org/citation.cfm?id=2975176&CFID=715860528&CFTOKEN=30536297>

摘要：

# Recently, distributed word embeddings trained by neural language models are commonly used for text classification with Convolutional Neural Networks (CNNs). In this paper, we propose a novel neural language model, Topic-based Skip-gram, to learn topic-based word embeddings for biomedical literature indexing with CNNs. Topic-based Skip-gram leverages textual content with topic models, e.g., Latent Dirichlet Allocation (LDA), to capture precise topic-based word relationship and then integrate it into distributed word embedding learning. We then describe two multimodal CNN architectures, which are able to employ different kinds of word embeddings at the same time for text classification. Through extensive experiments conducted on several real-world datasets, we demonstrate that combination of our Topic-based Skip-gram and multimodal CNN architectures outperforms state-of-the-art methods in biomedical literature indexing, clinical note annotation and general textual benchmark dataset classification.

# 最近，由神经语言模型训练的分布式字嵌入通常用于具有卷积神经网络（CNN）的文本分类。在本文中，我们提出一种新的神经语言模型，基于主题的Skip-gram，学习基于主题的词嵌入用于CNN的生物医学文献索引。基于主题的Skip-gram利用主题模型（例如，Latent Dirichlet分配（LDA））来利用文本内容来捕获精确的基于主题的词关系，然后将其集成到分布式词嵌入学习中。然后我们描述两个多模式CNN架构，其能够在文本分类的同时使用不同种类的字嵌入。通过对几个真实世界数据集进行广泛的实验，我们证明了基于主题的Skip-gram和多模态CNN架构的组合优于生物医学文献索引，临床注释注释和一般文本基准数据集分类中的最先进的方法。

# 关键词：文本分类，字嵌入，卷积神经网络，医学主题词

# 提出的问题：

# 在本文中，我们提出一种新的神经语言模型，基于主题的Skip-gram，学习基于主题的词嵌入用于CNN的生物医学文献索引。

# 解决方法：

# 基于主题的Skip-gram利用主题模型（例如，Latent Dirichlet分配（LDA））来利用文本内容来捕获精确的基于主题的词关系，然后将其集成到分布式词嵌入学习中。然后我们描述两个多模式CNN架构，其能够在文本分类的同时使用不同种类的字嵌入。

1. Bridging the Semantic Gaps of GPU Acceleration for Scale-out CNN-based Big Data Processing: Think Big, See Small

September 11 - 15, 2016

链接：

<http://dl.acm.org/citation.cfm?id=2967944&CFID=715860528&CFTOKEN=30536297>

摘要：

Convolutional Neural Networks (CNNs) have substantially advanced the state-of-the-art accuracies of object recognition, which is the core function of a myriad of modern multimedia processing techniques such as image/video processing, speech recognition, and natural language processing. GPU-based accelerators gained increasing attention because a large amount of highly parallel neurons in CNN naturally matches the GPU computation pattern. In this work, we perform comprehensive experiments to investigate the performance bottlenecks and overheads of current GPU acceleration platform for scale-out CNN-based big data processing.

In our characterization, we observe two significant semantic gaps: framework gap that lies between CNN-based data processing workflow and data processing manner in distributed framework; and the standalone gap that lies between the uneven computation loads at different CNN layers and fixed computing capacity provisioning of current GPU acceleration library. To bridge these gaps, we propose *D3NN*, a Distributed, Decoupled, and Dynamically tuned GPU acceleration framework for modern CNN architectures. In particular, *D3NN* features a novel analytical model that enables accurate time estimation of GPU accelerated CNN processing with only 5-10% error. Our evaluation results show the throughput of standalone processing node using *D3NN* gains up to 3.7X performance improvement over current standalone GPU acceleration platform. Our CNN-oriented GPU acceleration library with built-in dynamic batching scheme achieves up to 1.5X performance improvement over the non-batching scheme and outperforms the state-of-the-art deep learning library by up to 28% (performance mode) ~ 67% (memory-efficient mode).

卷积神经网络（CNN）已经基本上提高了对象识别的现有技术精度，这是诸如图像/视频处理，语音识别和自然语言处理的大量现代多媒体处理技术的核心功能。基于GPU的加速器获得越来越多的关注，因为CNN中的大量高度并行的神经元自然地匹配GPU计算模式。在这项工作中，我们进行全面的实验，以调查当前GPU加速平台的横向扩展基于CNN的大数据处理的性能瓶颈和开销。在我们的表征中，我们观察到两个重要的语义差距：框架间隙在基于CNN的数据处理工作流和分布式框架中的数据处理方式之间; 以及位于不同CNN层的不均衡计算负载和当前GPU加速库的固定计算能力供应之间的独立间隙。为了弥补这些缺陷，我们提出*ð 3 NN*，现代CNN架构分布式，解耦，并动态调整GPU加速的框架。特别是，*ð 3 NN*具有新颖的分析模型，使GPU的准确时间估计CNN加速处理，只有5-10％的误差。我们的评估结果表明，使用独立的处理节点的吞吐量*ð 3NN*增益高达3.7X性能改进过电流独立GPU加速平台。我们的基于CNN的GPU加速库与内置动态批处理方案相比，非批处理方案的性能提高了1.5倍，并且超越了最先进的深度学习库高达28％（性能模式）〜 67％（存储器高效模式）

**关键词：**分布式系统，深度学习，GPU，大数据

**提出的问题：**

我们进行全面的实验，以调查当前GPU加速平台的横向扩展基于CNN的大数据处理的性能瓶颈和开销。

**解决方法：**

在我们的表征中，我们观察到两个重要的语义差距：框架间隙在基于CNN的数据处理工作流和分布式框架中的数据处理方式之间; 以及位于不同CNN层的不均衡计算负载和当前GPU加速库的固定计算能力供应之间的独立间隙。为了弥补这些缺陷，我们提出*ð 3 NN*，现代CNN架构分布式，解耦，并动态调整GPU加速的框架。

1. Convolutional Matrix Factorization for Document Context-Aware Recommendation

September 15 - 19, 2016

链接：

<http://dl.acm.org/citation.cfm?id=2959165&CFID=715860528&CFTOKEN=30536297>

摘要：

Sparseness of user-to-item rating data is one of the major factors that deteriorate the quality of recommender system. To handle the sparsity problem, several recommendation techniques have been proposed that additionally consider auxiliary information to improve rating prediction accuracy. In particular, when rating data is sparse, document modeling-based approaches have improved the accuracy by additionally utilizing textual data such as reviews, abstracts, or synopses. However, due to the inherent limitation of the bag-of-words model, they have difficulties in effectively utilizing contextual information of the documents, which leads to shallow understanding of the documents. This paper proposes a novel context-aware recommendation model, convolutional matrix factorization (ConvMF) that integrates convolutional neural network (CNN) into probabilistic matrix factorization (PMF). Consequently, ConvMF captures contextual information of documents and further enhances the rating prediction accuracy. Our extensive evaluations on three real-world datasets show that ConvMF significantly outperforms the state-of-the-art recommendation models even when the rating data is extremely sparse. We also demonstrate that ConvMF successfully captures subtle contextual difference of a word in a document. Our implementation and datasets are available at <http://dm.postech.ac.kr/ConvMF>.

用户到项目评级数据的稀疏性是恶化推荐系统的质量的主要因素之一。为了处理稀疏问题，已经提出了几种推荐技术，其另外考虑辅助信息以提高评级预测精度。特别地，当评级数据稀疏时，基于文档建模的方法通过另外使用诸如评论，摘要或概要的文本数据来提高准确性。然而，由于词袋模型的固有限制，他们难以有效地利用文档的上下文信息，这导致对文档的浅的理解。本文提出一种新颖的上下文感知推荐模型，卷积矩阵分解（ConvMF），将卷积神经网络（CNN）集成到概率矩阵分解（PMF）。因此，ConvMF捕获文档的上下文信息，并进一步提高评级预测精度。我们对三个真实世界数据集的广泛评估表明，即使评级数据非常稀疏，ConvMF也显着优于最先进的推荐模型。我们还演示ConvMF成功捕获文档中单词的微妙上下文差异。我们的实现和数据集可以在<http://dm.postech.ac.kr/ConvMF>。

**关键词**：协同过滤，文件建模，推荐系统，深度学习，神经网络“环境感知推荐，contexual信息

**提出的问题：**

当评级数据稀疏时，基于文档建模的方法通过另外使用诸如评论，摘要或概要的文本数据来提高准确性。然而，由于词袋模型的固有限制，他们难以有效地利用文档的上下文信息，这导致对文档的浅的理解。

**解决方法：**

本文提出一种新颖的上下文感知推荐模型，卷积矩阵分解（ConvMF），将卷积神经网络（CNN）集成到概率矩阵分解（PMF）。因此，ConvMF捕获文档的上下文信息，并进一步提高评级预测精度。

1. MatConvNet: Convolutional Neural Networks for MATLAB

September 08 - 11, 2015

链接：

<http://dl.acm.org/citation.cfm?id=2807412&CFID=715860528&CFTOKEN=30536297>

摘要：

MatConvNet is an open source implementation of Convolutional Neural Networks (CNNs) with a deep integration in the MATLAB environment. The toolbox is designed with an emphasis on simplicity and flexibility. It exposes the building blocks of CNNs as easy-to-use MATLAB functions, providing routines for computing convolutions with filter banks, feature pooling, normalisation, and much more. MatConvNet can be easily extended, often using only MATLAB code, allowing fast prototyping of new CNN architectures. At the same time, it supports efficient computation on CPU and GPU, allowing to train complex models on large datasets such as ImageNet ILSVRC containing millions of training examples

MatConvNet是卷积神经网络（CNN）的开源实现，在MATLAB环境中具有深度集成。工具箱的设计注重简单性和灵活性。它将CNN的构建块暴露为易于使用的MATLAB函数，提供用于计算卷积的过滤器库，特性池，规范化等等的例程。MatConvNet可以轻松扩展，经常只使用MATLAB代码，允许快速原型的新的CNN架构。同时，它支持CPU和GPU的高效计算，允许在大型数据集上训练复杂模型，例如包含数百万个培训示例的ImageNet ILSVRC

**关键词：**计算机视觉，图像理解，机器学习，卷积神经网络，深度学习

**提出的问题：**

MatConvNet是卷积神经网络（CNN）的开源实现，在MATLAB环境中具有深度集成。工具箱的设计注重简单性和灵活性。

**解决方法：**

MatConvNet将CNN的构建块暴露为易于使用的MATLAB函数，提供用于计算卷积的过滤器库，特性池，规范化等等的例程。MatConvNet可以轻松扩展，经常只使用MATLAB代码，允许快速原型的新的CNN架构。

1. Learning Deep Features For MSR-bing Information Retrieval Challenge

September 08 - 11, 2015

链接：

<http://dl.acm.org/citation.cfm?id=2809928&CFID=715860528&CFTOKEN=30536297>

摘要：

Two tasks have been put forward in the MSR-bing Grand Challenge 2015. To address the information retrieval task, we raise and integrate a series of methods with visual features obtained by convolution neural network (CNN) models. In our experiments, we discover that the ranking strategies of Hierarchical clustering and PageRank methods are mutually complementary. Another task is fine-grained classification. In contrast to basic-level recognition, fine-grained classification aims to distinguish between different breeds or species or product models, and often requires distinctions that must be conditioned on the object pose for reliable identification. Current state-of-the-art techniques rely heavily upon the use of part annotations, while the bing datasets suffer both abundance of part annotations and dirty background. In this paper, we propose a CNN-based feature representation for visual recognition only using image-level information. Our CNN model is pre-trained on a collection of clean datasets and fine-tuned on the bing datasets. Furthermore, a multi-scale training strategy is adopted by simply resizing the input images into different scales and then merging the soft-max posteriors. We then implement our method into a unified visual recognition system on Microsoft cloud service. Finally, our solution achieved top performance in both tasks of the contest

在MSR-bing大挑战2015年提出了两个任务。为了解决信息检索任务，我们提出和集成一系列方法与卷积神经网络（CNN）模型获得的视觉特征。在我们的实验中，我们发现层次聚类和PageRank方法的排名策略是相互补充的。另一个任务是细粒度分类。与基本水平识别相反，细粒度分类旨在区分不同品种或物种或产品模型，并且通常需要区分，必须以物体姿势为条件进行可靠识别。当前最先进的技术严重依赖于部分注释的使用，而bing数据集遭受大量的部分注释和肮脏的背景。在本文中，我们提出一种基于CNN的特征表示，用于仅使用图像级信息的视觉识别。我们的CNN模型是对一组干净的数据集进行预训练，并对bing数据集进行微调。此外，通过简单地将输入图像大小调整为不同尺度，然后合并软最大后验，采用多尺度训练策略。然后，我们将我们的方法实现在Microsoft云服务上的统一的视觉识别系统。最后，我们的解决方案在比赛的两个任务中实现了最佳性能

**关键词：**视觉识别，信息检索

**提出的问题：**

当前最先进的技术严重依赖于部分注释的使用，而bing数据集遭受大量的部分注释和肮脏的背景。

**解决方法：**

在本文中，我们提出一种基于CNN的特征表示，用于仅使用图像级信息的视觉识别。我们的CNN模型是对一组干净的数据集进行预训练，并对bing数据集进行微调。

1. Convolutional neural networks for biomedical text classification: application in indexing biomedical articles

September 08 - 11, 2015

链接：

<http://dl.acm.org/citation.cfm?id=2808746&CFID=715860528&CFTOKEN=30536297>

摘要：

Building high accuracy text classifiers is an important task in biomedicine given the wealth of information hidden in unstructured narratives such as research articles and clinical documents. Due to large feature spaces, traditionally, discriminative approaches such as logistic regression and support vector machines with n-gram and semantic features (e.g., named entities) have been used for text classification where additional performance gains are typically made through feature selection and ensemble approaches. In this paper, we demonstrate that a more direct approach using convolutional neural networks (CNNs) outperforms several traditional approaches in biomedical text classification with the specific use-case of assigning medical subject headings (or MeSH terms) to biomedical articles. Trained annotators at the national library of medicine (NLM) assign on an average 13 codes to each biomedical article, thus semantically indexing scientific literature to support NLM's PubMed search system. Recent evidence suggests that effective automated efforts for MeSH term assignment start with binary classifiers for each term. In this paper, we use CNNs to build binary text classifiers and achieve an absolute improvement of over 3% in macro F-score over a set of selected hard-to-classify MeSH terms when compared with the best prior results on a public dataset. Additional experiments on 50 high frequency terms in the dataset also show improvements with CNNs. Our results indicate the strong potential of CNNs in biomedical text classification tasks.

构建高精度文本分类器是生物医学中的一个重要任务，因为隐藏在非结构化叙事（例如研究文章和临床文献）中的大量信息。由于大的特征空间，传统上，具有n元语法和语义特征（例如，命名实体）的逻辑回归和支持向量机的辨别方法已经用于文本分类，其中附加的性能增益通常通过特征选择和集合方法。在本文中，我们证明使用卷积神经网络（CNNs）更直接的方法胜过生物医学文本分类中的几种传统方法与具体用例的医疗主题标题（或MeSH术语）到生物医学文章。国家医学图书馆（NLM）的经过训练的注释者对每个生物医学文章平均分配13个代码，从而语义地索引科学文献以支持NLM的PubMed搜索系统。最近的证据表明有效的自动化努力为MeSH术语分配开始与每个术语的二进制分类器。在本文中，我们使用CNNs来构建二进制文本分类器，并且在一个选定的难以分类的MeSH术语集合中，与在公共数据集上的最佳在先结果相比，在宏观F评分中实现超过3％的绝对改善。在数据集中的50个高频项的附加实验也显示了CNN的改进。我们的研究结果表明CNNs在生物医学文本分类任务的巨大潜力。

**关键词：**文本分类，卷积神经网络，医学主题词

**提出的问题：**

国家医学图书馆（NLM）的经过训练的注释者对每个生物医学文章平均分配13个代码，从而语义地索引科学文献以支持NLM的PubMed搜索系统。最近的证据表明有效的自动化努力为MeSH术语分配开始与每个术语的二进制分类器。

**解决方法：**

在本文中，我们使用CNNs来构建二进制文本分类器，并且在一个选定的难以分类的MeSH术语集合中，与在公共数据集上的最佳在先结果相比，在宏观F评分中实现超过3％的绝对改善。

1. Automatic Text Document Summarization Based on Machine Learning

链接：<http://dl.acm.org/citation.cfm?id=2797099&CFID=715860528&CFTOKEN=30536297>

September 08 - 11, 2015

摘要：

The need for automatic generation of summaries gained importance with the unprecedented volume of information available in the Internet. Automatic systems based on extractive summarization techniques select the most significant sentences of one or more texts to generate a summary. This article makes use of Machine Learning techniques to assess the quality of the twenty most referenced strategies used in extractive summarization, integrating them in a tool. Quantitative and qualitative aspects were considered in such assessment demonstrating the validity of the proposed scheme. The experiments were performed on the CNN-corpus, possibly the largest and most suitable test corpus today for benchmarking extractive summarization strategies.

自动生成摘要的需要随着互联网上可用的前所未有的信息量变得越来越重要。基于提取概括技术的自动系统选择一个或多个文本的最重要的句子以产生概要。本文利用机器学习技术来评估抽取式摘要中使用的二十个最引用的策略的质量，将它们集成到一个工具中。在此类评估中考虑了定量和定性方面，证明了拟议方案的有效性。对CNN语料库进行实验，可能是目前最大和最合适的测试语料库，用于基准抽取汇总策略。

**关键词**：文本摘要，采掘特点，句子评分方法

**提出的问题：**

自动生成摘要的需要随着互联网上可用的前所未有的信息量变得越来越重要。基于提取概括技术的自动系统选择一个或多个文本的最重要的句子以产生概要。

**解决方法：**

本文利用机器学习技术来评估抽取式摘要中使用的二十个最引用的策略的质量，将它们集成到一个工具中。