**1. Semantic proximity search on graphs with metagraph-based learning**

**基于段落学习的图形上的语义接近搜索**

Given ubiquitous graph data such as the Web and social networks, proximity search on graphs has been an active research topic. The task boils down to measuring the proximity between two nodes on a graph. Although most earlier studies deal with homogeneous or bipartite graphs only, many real-world graphs are heterogeneous with objects of various types, giving rise to different semantic classes of proximity. For instance, on a social network two users can be close for different reasons, such as being classmates or family members, which represent two distinct classes of proximity. Thus, it becomes inadequate to only measure a “generic” form of proximity as previous works have focused on. In this paper, we identify metagraphs as a novel and effective means to characterize the common structures for a desired class of proximity. Subsequently, we propose a family of metagraph-based proximity, and employ a supervised technique to automatically learn the right form of proximity within its family to suit the desired class. As it is expensive to match (i.e., find the instances of) a metagraph, we propose the novel approaches of dual-stage training and symmetry-based matching to speed up. Finally, our experiments reveal that our approach is significantly more accurate and efficient. For accuracy, we outperform the baselines by 11% and 16% in NDCG and MAP, respectively. For efficiency, dual-stage training reduces the overall matching cost by 83%, and symmetry-based matching further decreases the cost of individual metagraphs by 52%.

给定无处不在的图形数据，如Web和社交网络，图形上的邻近搜索已经成为一个积极的研究主题。任务归结为测量图上两个节点之间的接近度。虽然大多数早期的研究只处理同质或二分图，许多现实世界的图形是与各种类型的对象的异构，产生不同的接近的语义类。例如，在社交网络上，两个用户可以由于不同的原因而关闭，诸如作为同学或家庭成员，其代表两个不同的接近类别。因此，它变得不足以仅测量如先前作品所关注的“通用”形式的接近度。**在本文中，我们确定段落作为一种新颖的和有效的手段来表征所需类别的邻近的共同结构。随后，我们提出一个基于段落邻近的家庭，并采用监督技术，以自动学习在其家庭内适合所需类的正确形式的接近**。由于匹配（即，找到实例）段是昂贵的，我们提出了双阶段训练和基于对称的匹配的新颖方法以加速。最后，我们的实验表明，我们的方法是更准确和有效率。为了准确性，我们分别在NDCG和MAP中分别比基线高11％和16％。为了效率，双阶段训练将总体匹配成本降低了83％，基于对称的匹配进一步将单个段的成本降低了52％。

**2. Context-aware advertisement recommendation for high-speed social news feeding**

**上下文感知广告推荐高速社会新闻喂养**

Social media advertising is a multi-billion dollar market and has become the major revenue source for Facebook and Twitter. To deliver ads to potentially interested users, these social network platforms learn a prediction model for each user based on their personal interests. However, as user interests often evolve slowly, the user may end up receiving repetitive ads. In this paper, we propose a context-aware advertising framework that takes into account the relatively static personal interests as well as the dynamic news feed from friends to drive growth in the ad click-through rate. To meet the real-time requirement, we first propose an online retrieval strategy that finds k most relevant ads matching the dynamic context when a read operation is triggered. To avoid frequent retrieval when the context varies little, we propose a safe region method to quickly determine whether the top-k ads of a user are changed. Finally, we propose a hybrid model to combine the merits of both methods by analyzing the dynamism of news feed to determine an appropriate retrieval strategy. Extensive experiments conducted on multiple real social networks and ad datasets verified the efficiency and robustness of our hybrid model.

社交媒体广告是一个数十亿美元的市场，已成为Facebook和Twitter的主要收入来源。为了向潜在感兴趣的用户递送广告，这些社交网络平台基于每个用户的个人兴趣来学习每个用户的预测模型。然而，由于用户兴趣经常变慢，用户可能最终接收到重复的广告。**在本文中，我们提出一个上下文感知广告框架，考虑到相对静态的个人兴趣以及来自朋友的动态新闻Feed，以推动广告点击率的增长。为了满足实时要求，我们首先提出在线检索策略，其在触发读取操作时找到与动态上下文匹配的k个最相关的广告。**为了避免在上下文变化不大时频繁检索，我们提出了一种安全区域方法，以快速确定用户的前k个广告是否被更改。最后，我们提出一种混合模型，通过分析新闻馈送的动态性，以确定适当的检索策略，结合两种方法的优点。在多个真实社交网络和广告数据集上进行的广泛实验验证了我们的混合模型的效率和鲁棒性。

**3. A Model-based Approach for Text Clustering with Outlier Detection**

**基于模型的方法用于文本聚类与异常值检测**

Text clustering is a challenging problem due to the high-dimensional and large-volume characteristics of text datasets. In this paper, we propose a collapsed Gibbs Sampling algorithm for the Dirichlet Process Multinomial Mixture model for text clustering (abbr. to GSDPMM) which does not need to specify the number of clusters in advance and can cope with the high-dimensional problem of text clustering. Our extensive experimental study shows that GSDPMM can achieve significantly better performance than three other clustering methods and can achieve high consistency on both long and short text datasets. We found that GSDPMM has low time and space complexity and can scale well with huge text datasets. We also propose some novel and effective methods to detect the outliers in the dataset and obtain the representative words of each cluster.

由于文本数据集的高维度和大体积特征，文本聚类是一个具有挑战性的问题。 **在本文中，我们提出了一种用于文本聚类（缩写为GSDPMM）的Dirichlet过程多项混合模型的折叠吉布斯采样算法，其不需要预先指定聚类的数量，并且可以应对文本的高维问题 聚类。** 我们广泛的实验研究表明，GSDPMM可以实现明显比其他三种聚类方法更好的性能，并可以实现长文本和短文本数据集的高一致性。 我们发现GSDPMM具有较低的时间和空间复杂性，可以很好地扩展与巨大的文本数据集。 我们还提出一些新的和有效的方法来检测数据集中的异常值，并获得每个聚类的代表性词。

**4. OptImatch: Semantic System for Query Problem Determination**

**用于查询问题确定的语义系统**

Query performance problem determination is usually performed by analyzing query execution plans (QEPs). Analyzing complex QEPs is excessively time consuming and existing automatic problem determination tools do not provide ability to perform analysis with flexible user-defined problem patterns. We present the novel OptImatch system that allows a relatively naive user to search for patterns in QEPs and get recommendations from an expert and user customizable knowledge base. Our system transforms a QEP into an RDF graph. We provide a web graphical interface for the user to describe a pattern that is transformed with handlers into a SPARQL query. The SPARQL query is matched against the abstracted RDF graph and any matched parts of the graph are relayed back to the user. With the knowledge base the system automatically matches stored patterns to the QEPs by adapting dynamic context through developed tagging language and ranks recommendations using statistical correlation analysis.

查询性能问题确定通常通过分析查询执行计划（QEP）来执行。 分析复杂的QEP是非常耗时的，并且现有的自动问题确定工具不提供利用灵活的用户定义的问题模式执行分析的能力。 我们提出了新的OptImatch系统，允许一个相对幼稚的用户搜索QEP中的模式，并从专家和用户可定制的知识库获得建议。 我们的系统将QEP转换为RDF图。 我们提供了一个Web图形界面，供用户描述使用处理程序转换为SPARQL查询的模式。 SPARQL查询与抽象的RDF图进行匹配，并且图的任何匹配部分被中继回用户。 **利用知识库，系统通过使用统计相关分析通过开发的标记语言适配动态上下文来自动地将存储的模式匹配到QEP，并且推荐建议。**

**5. Semantic-Aware Blocking for Entity Resolution (Extended abstract)**

**用于实体解析的语义感知阻塞**

n this paper, we propose a semantic-aware blocking framework for entity resolution (ER). The proposed framework is built using locality-sensitive hashing (LSH) techniques, which efficiently unifies both textual and semantic features into an ER blocking process. In order to understand how similarity metrics may affect the effectiveness of ER blocking, we study the robustness of similarity metrics and their properties in terms of LSH families. Then, we present how the semantic similarity of records can be captured, measured, and integrated with LSH techniques over multiple similarity spaces. In doing so, the proposed framework can support efficient similarity searches on records in both textual and semantic similarity spaces, yielding ER blocking with improved quality. We have evaluated the proposed framework over two real-world data sets, and compared it with the state-of-the-art blocking techniques. Our experimental study shows that the combination of semantic similarity and textual similarity can considerably improve the quality of blocking. Furthermore, due to the probabilistic nature of LSH, this semantic-aware blocking framework enables us to build fast and reliable blocking for performing entity resolution tasks in a large-scale data environment.

在本文中，我们提出了用于实体解析（ER）的语义感知阻塞框架。所提出的框架使用局部敏感散列（LSH）技术来构建，其有效地将文本和语义特征统一为ER阻塞过程。为了了解相似性度量如何影响ER阻断的有效性，我们研究相似性度量的鲁棒性及其在LSH家族方面的性质。然后，我们展示了如何在多个相似空间中捕获，测量和与LSH技术集成的记录的语义相似性。在这样做时，所提出的框架可以支持对文本和语义相似性空间中的记录的高效的相似性搜索，产生具有改进的质量的ER阻塞。我们已经评估了两个真实世界数据集的框架，并将其与最先进的阻塞技术进行了比较。我们的实验研究表明，语义相似性和文本相似性的组合可以显着提高阻塞的质量。此外，由于LSH的概率性质，这种语义感知阻塞框架使我们能够建立快速和可靠的阻塞，用于在大规模数据环境中执行实体解析任务。

**6. Linking Tweets to News: A Framework to Enrich Short Text Data in Social Media（2013）**

**将Tweets链接到新闻：在社交媒体中丰富短文本数据的框架**

Many current Natural Language Processing [NLP] techniques work well assuming a large context of text as input data. However they become ineffective when applied to short texts such as Twitter feeds. To overcome the issue, we want to find a related newswire document to a given tweet to provide contextual support for NLP tasks. This requires robust modeling and understanding of the semantics of short texts. The contribution of the paper is two-fold: 1. we introduce the Linking-Tweets-to-News task as well as a dataset of linked tweet-news pairs, which can benefit many NLP applications; 2. in contrast to previous research which focuses on lexical features within the short texts (text-to-word information), we propose a graph based latent variable model that models the inter short text correlations (text-to-text information). This is motivated by the observation that a tweet usually only covers one aspect of an event. We show that using tweet specific feature (hashtag) and news specific feature (named entities) as well as temporal constraints, we are able to extract text-to-text correlations, and thus completes the semantic picture of a short text. Our experiments show significant improvement of our new model over baselines with three evaluation metrics in the new task.

许多当前的自然语言处理（NLP）技术工作良好，假设文本作为输入数据的大上下文。然而，当应用于诸如Twitter馈送的短文本时，它们变得无效。为了克服这个问题，我们希望找到一个相关的newswire文档给一个给定的tweet，**以提供NLP任务的上下文支持**。这需要对短文本的语义的鲁棒建模和理解。本文的贡献是双重的：1.我们介绍Linking-Tweets-to-News任务以及链接的tweet-news对的数据集，这可以使许多NLP应用程序受益; 2.与以前的研究集中于短文本（文本到词信息）中的词汇特征相反，我们提出了一个基于图的潜变量模型，其模型的短文本相关性（文本到文本信息）。这是由于观察发现推文通常只涵盖事件的一个方面。我们显示使用tweet特定特征（hashtag）和新闻特定特征（命名实体）以及时间约束，我们能够提取文本到文本的相关性，从而完成短文本的语义图片。我们的实验显示我们的新模型相对于基线的显着改进，在新任务中有三个评估指标。

2015年

**7. Short Text Understanding Through Lexical-Semantic Analysis**

**短语通过词汇语义分析理解**

Understanding short texts is crucial to many applications, but challenges abound. First, short texts do not always observe the syntax of a written language. As a result, traditional natural language processing methods cannot be easily applied. Second, short texts usually do not contain suffi cient statistical signals to support many state-of-the-art approaches for text processing such as topic modeling. Third, short texts are usually more ambiguous. We argue that knowledge is needed in order to better understand short texts. In this work, we use lexicalsemantic knowledge provided by a well-known semantic network for short text understanding. Our knowledge-intensive approach disrupts traditional methods for tasks such as text segmentation, part-of-speech tagging, and concept labeling, in the sense that we focus on semantics in all these tasks. We conduct a comprehensive performance evaluation on real-life data. The results show that knowledge is indispensable for short text understanding, and our knowledge-intensive approaches are effective in harvesting semantics of short texts.

理解短文本对于许多应用程序至关重要，但挑战众多。首先，短文本不总是遵守书面语言的语法。因此，传统的自然语言处理方法不能容易地应用。第二，短文本通常不包含足够的统计信号以支持诸如主题建模的文本处理的许多现有技术方法。第三，短文本通常更加模糊。我们认为，为了更好地理解短文本，需要知识。在这项工作中，我们**使用由着名的语义网络提供的词汇语言知识来进行短文本理解**。我们的知识密集型方法破坏了传统的任务方法，例如文本分段，词性标注和概念标注，我们专注于所有这些任务中的语义。我们对实际数据进行全面的性能评估。结果表明，知识是短文本理解必不可少的，我们的知识密集型方法有效收获短文本的语义。

**8. Approximate keyword search in semantic trajectory database**

**语义轨迹数据库中的近似关键词搜索**

Driven by the advances in location positioning techniques and the popularity of location sharing services, semantic enriched trajectory data have become unprecedentedly available. While finding relevant Point-of-Interest (POIs) based on users' locations and query keywords has been extensively studied in the past years, it is largely untouched to explore the keyword queries in the context of semantic trajectory database. In this paper, we study the problem of approximate keyword search in massive semantic trajectories. Given a set of query keywords, an approximate keyword query of semantic trajectory (AKQST) returns k trajectories that contain the most relevant keywords to the query and yield the least travel effort in the meantime. The main difference between AKQST and conventional spatial keyword queries is that there is no query location in AKQST, which means the search area cannot be localized. To capture the travel effort in the context of query keywords, a novel utility function, called spatio-textual utility function, is first defined. Then we develop a hybrid index structure called GiKi to organize the trajectories hierarchically, which enables pruning the search space by spatial and textual similarity simultaneously. Finally an efficient search algorithm and fast evaluation of the minimum value of spatio-textual utility function are proposed. The results of our empirical studies based on real check-in datasets demonstrate that our proposed index and algorithms can achieve good scalability.

在位置定位技术的进步和位置共享服务的普及的驱动下，语义丰富的轨迹数据已经变得前所未有地可用。虽然在过去几年中已经广泛研究基于用户的位置和查询关键词找到相关的兴趣点（POI），但在语义轨迹数据库的语境中探索关键字查询是很大程度上没有改变的。在本文中，我们**研究大规模语义轨迹中的近似关键词搜索的问题**。**给定一组查询关键词，语义轨迹（AKQST）的近似关键词查询返回k个轨迹，其包含与查询最相关的关键词，并且在此期间产生最少的旅行努力。** AKQST和常规空间关键字查询之间的主要区别是在AKQST中没有查询位置，这意味着搜索区域不能被本地化。为了在查询关键词的上下文中捕获旅行努力，首先定义称为空间文本效用函数的新颖效用函数。然后我们开发一个称为GiKi的混合索引结构来分层组织轨迹，这使得通过空间和文本相似性同时修剪搜索空间。最后提出了一种高效的搜索算法和快速评价空间文本效用函数的最小值。基于实际检入数据集的实证研究的结果表明，我们提出的索引和算法可以实现良好的可扩展性。

**9. PandaSearch: A fine-grained academic search engine for research documents**

**PandaSearch：用于研究文档的细粒度的学术搜索引擎**

In the world of academia, research documents enable the sharing and dissemination of scientific discoveries. During these “big data” times, academic search engines are widely used to find the relevant research documents. Considering the domain of computer science, a researcher often inputs a query with a specific goal to find an algorithm or a theorem. However, to this date, the return result of most search engines is just as a list of related papers. Users have to browse the results, download the interesting papers and look for the desired information, which is obviously laborious and inefficient. In this paper, we present a novel academic search system, called PandaSearch, that returns the results with a fine-grained interface, where the results are well organized by different categories, such as definitions, theorems, lemmas, algorithms and figures. The key technical challenges in our system include the automatic identification and extraction of different parts in a research document, the discovery of the main topic phrases for a definition or a theorem, and the recommendation of related definitions or figures to elegantly satisfy the search intention of users. Based on this, we have built a user friendly search interface for users to conveniently explore the documents, and find the relevant information.

在学术界，研究文件能够分享和传播科学发现。在这些“大数据”时代，学术搜索引擎被广泛用于寻找相关的研究文献。考虑到计算机科学的领域，研究人员通常输入具有特定目标的查询以找到算法或定理。然而，到目前为止，大多数搜索引擎的返回结果只是作为相关论文的列表。用户必须浏览结果，下载有趣的文件并寻找所需的信息，这显然是费力和低效的。在本文中，我们提出一个新的**学术搜索系统**，称为PandaSearch，返回结果与细粒度的接口，其中的结果是由不同的类别，如定义，定理，引理，算法和数字组织良好。我们系统中的关键技术挑战包括在研究文档中自动识别和提取不同部分，发现用于定义或定理的主要主题短语，以及相关定义或图形的建议，以优雅地满足搜索意图用户。基于此，我们建立了一个用户友好的搜索界面，方便用户浏览文档，并找到相关信息。

**10. CROWN: A Context-aware RecOmmender for Web News**

**CROWN：Web新闻的上下文感知RecOmmender**

It is popular for most people to read news online since the web sites can provide access to news articles from millions of sources around the world. For these news web sites, the key challenge is to help users find related news articles to read. In this paper, we present a system called CROWN (Context-aware RecOmmender for Web News) to do Chinese news recommendation. By recommendation, the system can retrieve personalized fresh and relevant news articles to mobile users according to their particular context. Differing from existing mobile news applications which employ rather simple strategies for news recommendation, CROWN integrates the contextual information in prediction by modeling the data as a tensor. Such context information usually includes the time, the location, etc. This demo paper presents the implementation of the whole procedure of news recommendation in the system of CROWN. Experimental results on a large corpus of newly-published Chinese web news show its performance is satisfactory

它是受欢迎的大多数人阅读新闻在线，因为网站可以提供访问来自世界各地的数百万来源的新闻文章。对于这些新闻网站，主要的挑战是帮助用户找到相关的新闻文章阅读。在本文中，我们提出一个系统，**称为CROWN（上下文感知RecOmmender for Web新闻）做中文新闻推荐。通过推荐，系统可以根据移动用户的特定上下文来检索个性化的新鲜和相关的新闻文章。**与现有的移动新闻应用程序不同，其采用相当简单的策略来进行新闻推荐，CROWN通过将数据建模为张量来将上下文信息集成到预测中。这样的上下文信息通常包括时间，位置等。本演示文稿介绍了CROWN系统中新闻推荐的整个过程的实现。新出版的中文网新闻的大型实验结果表明，其性能令人满意

**11. LearningAssistant: A novel learning resource recommendation system**

**LearningAssistant：一种新颖的学习资源推荐系统**

Reading online content for educational, learning, training or recreational purposes has become a very popular activity. While reading, people may have difficulty understanding a passage or wish to learn more about the topics covered by it, hence they may naturally seek additional or supplementary resources for the particular passage. These resources should be close to the passage both in terms of the subject matter and the reading level. However, using a search engine to find such resources interrupts the reading flow. It is also an inefficient, trial-and-error process because existing web search and recommendation systems do not support large queries, they do not understand semantic topics, and they do not take into account the reading level of the original document a person is reading. In this demo, we present LearningAssistant, a novel system that enables online reading material to be smoothly enriched with additional resources that can supplement or explain any passage from the original material for a reader on demand. The system facilitates the learning process by recommending learning resources (documents, videos, etc) for selected text passages of any length. The recommended resources are ranked based on two criteria (a) how they match the different topics covered within the selected passage, and (b) the reading level of the original text where the selected passage comes from. User feedback from students who use our system in two real pilots, one with a high school and one with a university, for their courses suggest that our system is promising and effective.

阅读用于教育，学习，培训或娱乐目的的在线内容已经成为非常受欢迎的活动。在阅读时，人们可能难以理解一段话或希望更多地了解它所涉及的主题，因此他们可能自然地为特定段落寻求额外的或补充的资源。这些资源应该接近主题和阅读水平的通道。然而，使用搜索引擎来找到这样的资源中断读取流。这也是一个效率低下的试错法过程，因为现有的网络搜索和推荐系统不支持大型查询，他们不理解语义主题，并且他们不考虑一个人正在阅读的原始文档的阅读水平。在这个演示中，我们提出了LearningAssistant，一个新的系统，使在线阅读材料能够平滑地丰富与额外的资源，可以补充或解释从原始材料的任何通道的读者请求。**该系统通过为任何长度的所选文本段推荐学习资源（文档，视频等）来促进学习过程**。推荐的资源基于两个标准（a）它们如何匹配所选段落中涵盖的不同主题，以及（b）所选段落所来自的原始文本的阅读水平来排名。在两个真正的飞行员，一个高中和一个大学使用我们的系统的学生的用户反馈，他们的课程表明我们的系统是有希望和有效的。

**12. Inferencing in information extraction: Techniques and applications**

**信息提取中的参考：技术和应用**

Information extraction at Web scale has become one of the most important research topics in data management since major commercial search engines started incorporating knowledge in their search results a couple of years ago [1]. Users increasingly expect structured knowledge as answers to their search needs. Using Bing as an example, the result page for “Lionel Messi” is full of structured knowledge facts, such as his birthday and awards. The research efforts towards improving the accuracy and coverage of such knowledge bases have led to significant advances in Information Extraction techniques [2], [3]. As the initial challenge of accurately extracting facts for popular entities are being addressed, more difficult challenges have emerged such as extending knowledge coverage to long tail entities and domains, understanding interestingness and usefulness of facts within a given context, and addressing information-seeking needs more directly and accurately. In this tutorial, we will survey the recent research efforts and provide an introduction to the techniques that address those challenges, and the applications that benefit from the adoption of those techniques. In particular, this tutorial will focus on a variety of techniques that can be broadly viewed as knowledge inferencing, i.e., combining multiple data sources and extraction techniques to verify existing knowledge and derive new knowledge. More specifically, we focus on four main categories of inferencing techniques: 1) deep natural language processing using machine learning techniques, 2) data cleaning using integrity constraints, 3) large-scale probabilistic reasoning, and 4) leveraging human expertise for domain knowledge extraction.

Web规模的信息提取已经成为数据管理中最重要的研究课题之一，因为主要的商业搜索引擎在几年前开始在他们的搜索结果中引入知识[1]。用户越来越期望结构化知识作为其搜索需求的答案。使用Bing作为例子，“Lionel Messi”的结果页面中充满了结构化的知识事实，例如他的生日和奖励。为提高这些知识库的准确性和覆盖面的研究努力已经导致信息提取技术的显着进步[2]，[3]。由于正在解决为大众实体准确提取事实的初始挑战，出现了更为困难的挑战，例如将知识覆盖扩展到长尾实体和领域，了解特定背景下事实的兴趣和有用性，以及满足信息寻求需求直接和准确。在本教程中，我们将调查最近的研究工作，并介绍解决这些挑战的技术，以及从采用这些技术中受益的应用。特别地，本教程将集中于可以被广泛地视为知识推断的各种技术，即组合多个数据源和提取技术以验证现有知识并导出新知识。更具体地说，我们关注四种主要的推理技术类型：**1）使用机器学习技术的深度自然语言处理，2）使用完整性约束的数据清理，3）大规模概率推理，以及4）利用人类专业知识来进行领域知识提取。**

**13. Goals in Social Media, information retrieval and intelligent agents**

**社交媒体，信息检索和智能代理的目标**

This tutorial provides a comprehensive and cohesive overview of goal modeling and recognition approaches by the Information Retrieval, the Artificial Intelligence and the Social Media communities. We will examine how these fields restrict the domain of study and how they capture notions easily perceived by humans' intuition but difficult to be formally defined and handled algorithmically. It is the purpose of this tutorial to provide a solid framework for placing existing work into perspective and highlight critical open challenges that will act as a springboard for researchers and practitioners in database systems, social data, and the Web, as well as developers of web-based, database-driven, and social applications, to work towards more user-centric systems and applications.

本教程为信息检索，人工智能和社交媒体社区提供了目标建模和识别方法的全面和一致的概述。 我们将研究这些领域如何限制研究领域，以及它们如何捕捉人类直觉容易感知到的概念，但难以正式定义和处理算法。 本教程的目的是提供一个坚实的框架，将现有的工作置于角度，突出重要的开放性挑战，将作为数据库系统，社交数据和Web中的研究人员和从业者以及Web开发人员的跳板 基于数据库驱动和社交应用程序，致力于更以用户为中心的系统和应用程序。