FacetGist: Collective Extraction of Document Facets in Large Technical Corpora

Tarique Siddiqui* Xiang Ren* Aditya Parameswaran Jiawei Han
University of Illinois at Urbana-Champaign, Urbana, IL, USA
{tsiddiq2, xren7, adityagp, hanj}@illinois.edu

ABSTRACT

Given the large volume of technical documents available, it is crucial to automatically organize and categorize these documents to be able to understand and extract value from them. Towards this end, we introduce a new research problem called Facet Extraction. Given a collection of technical documents, the goal of Facet Extraction is to automatically label each document with a set of concepts for the key facets (*e.g.*, application, technique, evaluation metrics, and dataset) that people may be interested in. Facet Extraction has numerous applications, including document summarization, literature search, patent search and business intelligence. The major challenge in performing Facet Extraction arises from multiple sources: concept extraction, concept to facet matching, and facet disambiguation. To tackle these challenges, we develop FacetGist, a framework for facet extraction. Facet Extraction involves constructing a graph-based heterogeneous network to capture information available across multiple *local* sentence-level features, as well as *global* context features. We then formulate a joint optimization problem, and propose an efficient algorithm for graph-based label propagation to estimate the facet of each concept mention. Experimental results on technical corpora from two domains demonstrate that Facet Extraction can lead to an improvement of over 25% in both precision and recall over competing schemes.

1. INTRODUCTION

With the ever-increasing number of technical documents being generated every day, including, but not limited to, patent folios, legal cases, real-estate agreements, historical archives, and scientific literature, there is a crucial need to develop automation that can identify the *concepts* for *key facets* for each document, so that readers can quickly get a sense for what the document is about, or search and retrieve documents based on these facets. Consider the domain of scientific publications, one we are all intimately familiar with. Given a new scientific paper, it is impossible for a reader to instantly understand the *techniques* being used, the kinds of *applications* that are addressed, or the *metrics* that are used to ascertain whether the techniques have good performance. Thus, we pose the following question: *Can we develop algorithms that can efficiently and automatically identify the key facets of each document in a large technical document corpora, with little manual supervision?*

Technique	Application
conditional random field; unsupervised learning; support vector machine; hidden markov model	Document summarization; sequence labeling; statistical classification
Evaluation Metric	**Dataset**
F1; Rouge-2	DUC

Table 1: Example of extracted facets for a publication: Document summarization using conditional random fields [26]

One may be tempted to consider using vanilla Natural Language Processing techniques to extract these facets from each document; however, since there is a vast difference in the language used within these technical documents as opposed to ordinary text, training domain-specific techniques would require large volumes of labeled training data. Yet another approach is to use citation networks, which would apply to all of the technical document types described above, and has been used for document retrieval and recommendation (*e.g.*, literature search [4], citation recommendation [25]), citation analysis (*e.g.*, paper and author impact study [23, 1]), and community discovery [9]. However, the citation networks completely ignore the textual data within the technical document, and hence can only provide a superficial understanding of the contents of the document.

Therefore, we identify a novel research problem, called Facet Extraction, in making sense of a large corpus of technical documents: given a collection of documents, a set of target facets users are interested in, and a small number of seed examples of concepts for each facet, the task of Facet Extraction is to label each document with a ranked list of concepts for each facet. The result of Facet Extraction is, thus, a summary of the major information of each document into a structured, multi-dimensional representation format, where the target facets serve as different attributes, and extracted concepts correspond to the attribute values (see Table 1).

Extracted facets largely enrich the original structured bibliographic meta information (*e.g.*, authors, venues, keywords), and thus enables a wide range of interesting applications. For example, in a literature search, facets can be used to answer questions such as "which techniques are used in this paper?" and "what are the applications of this work?" (see Table 1), which require a deeper understanding of the paper semantics than analyzing the author-generated keyword list. One can also answer questions like "what are the popular applications in the Natural Language Processing or the

*Equal contribution.

CIKM'16 , October 24-28, 2016, Indianapolis, IN, USA
 ISBN 978-1-4503-4073-1/16/10...$15.00
DOI: http://dx.doi.org/10.1145/2983323.2983828

Database Systems community?" (see Table 8) and "how does the facet of *entity recognition* vary across different communities?" (see Fig. 3), by aggregating the facets statistics across the database. Such results enable the discovery of ideas and the dynamics of a research topic or community in an effective and efficient way.

Challenges. The task of Facet Extraction involves several research challenges that arise because of the complexity and domain specific nature of technical documents:

• **Domain specific Concept Extraction:** The candidate concepts used for identifying key facets must not only be syntactically correct based on *local* sentence clues (*e.g.*, part-of-speech (POS) tag patterns, suffix or prefix and capitalization), but also must be statistically significant to the document based on *global* information (*e.g.*, intra-document frequency, corpus-level popularity). Existing weakly-supervised methods for entity and concept recognition [11, 31, 10] are based on noun phrase chunking and supervised entity recognition that rely on only local syntactical features, and are trained over general-domain corpora (*e.g.*, news articles). Therefore, the candidates generated using these methods are not always representative of the contents of technical documents. Alternatively, these methods require substantial additional training cost to work on technical corpora.

• **Concept to Facet Matching:** Existing work on keyphrase extraction [13, 32] and phrase mining [17, 21] are concerned with extracting words and multi-word sequences from the text without identifying their potential facets. Document summarization and topic modeling methods [26, 5] identify the hidden thematic structure of documents by highlighting key sentences and clusters of words that co-occur frequently. While complementary to our task, these methods do not provide the level of detail or granularity needed for differentiating concepts belonging to various facets such as techniques, applications, datasets, or evaluation metrics. Consider for example the phrase "used decision tree for classification" — here, decision tree is from the facet 'technique' while classification belongs to the facet 'application'. However, both document summarization and topic modeling methods would cluster these two concepts together as they co-occur quite frequently.

• **Facet Disambiguation:** Many concepts change facets across documents. For example, while "*sequential pattern mining*" is a *technique* in a phrase mining paper [17], it is an *application* for another paper [2]. In our evaluation dataset (a subset of documents from DBLP and ACL labeled by human experts), over 25% of concepts are found to represent different facets across different documents. Unfortunately, existing weakly-supervised entity typing methods assume that each concept is associated with a single type throughout the corpus, and thus cannot distinguish their facets when their role varies across documents.

Solution outline. Our proposed method leverages several intuitive ideas to address these aforementioned challenges. First, to conduct effective domain specific concept extraction, we consider both local and global text clues when extracting significant and accurate phrases as concept candidates. We integrate a domain-agnostic phrase segmentation method with part-of-speech tag constraints and interestingness-based filtering, all of which require minimal linguistic assumptions on the document collection. Second, to address concept to facet matching and facet ambiguity, we model the facet of a concept in different documents separately by studying its sentence-level signals as well as document-level and corpus-level structure information (*i.e.*, section structure, and document topics), and integrate these signals in a heterogeneous network.

To systematically integrate these ideas, we propose a novel Facet Extraction framework — FacetGist. First, we perform document logical structure parsing and latent Dirichlet allocation to obtain the section structure and a list of topics for each document, respectively. Second, a domain-agnostic phrase segmentation algorithm is applied to the document collection to generate quality phrases, and then, we filter out extraneous phrases using POS tag patterns and an interestingness measure to obtain quality concept candidates and their surrounding relation phrases simultaneously. Third, a heterogeneous graph is constructed to integrate both local and global context information for the candidate concepts. Document topics are bound with candidate concepts and sections to disambiguate their facets. Co-occurrences between concepts, relation phrases, suffixes, and sections are modeled as different relations in the graph. Finally, a joint optimization problem is formulated following graph-based label propagation, that estimates the confidence on how likely each concept serves as the target facets for the document.

Contributions. The major novel contributions of this paper are summarized as follows.

1. We define and study a novel task in the domain of technical documents understanding, Facet Extraction, which aims to highlight the key facets of a technical document using accurate and relevant concepts (Section 2).
2. We design a data-driven concept extraction method that uses a distantly-supervised phrase segmentation algorithm, POS tag patterns, and an interestingness measure for concepts (Section 3).
3. We propose a framework which integrates both *local text signals* (*e.g.*, relation phrases, concept suffix) with *global structure signals* (*e.g.*, paper sections, topics) by constructing a heterogeneous graph, and formulate a joint optimization problem for estimating the facets of all the candidate concepts collectively (Section 4).
4. We conduct experiments for both quantitative and qualitative evaluation of our approach on real world datasets (DBLP and ACL). Experimental results depict a significant improvement in both candidate extraction as well as facet identification as compared to the existing techniques (Section 5).

2. BACKGROUND AND PROBLEM

Concept: A *concept* (as defined in [21]) is a single word or a multi-word phrase that represents a real or imaginary entity or idea that many users may be interested in (*i.e.*, is significant in the corpus), and does not contain any extraneous words such that excluding them would identify the same entity (*i.e.*, is concise). Let $\mathcal{C} = \{c_1, c_2, \ldots, c_{|\mathcal{C}|}\}$ denote the set of concepts that occur in a collection of documents $\mathcal{D} = \{d_1, \ldots, d_D\}$ where d_i is a document. A *concept mention*, $m = d\_c$, represents the occurrence of a concept $c \in \mathcal{C}$ in a specific document $d \in \mathcal{D}$. Note that multiple concepts can refer to the same entity (*e.g.*, both "*SVM*" and "*support vector machine*" refer to the real-world entity *Support Vector Machine*).

Facet: Users often search, read and compare technical documents in terms of different facets (*i.e.*, specific ways that a document can be considered), *e.g.*, techniques, applications, evaluation metrics, or datasets. A concept mentioned in a document can be categorized based on these facets

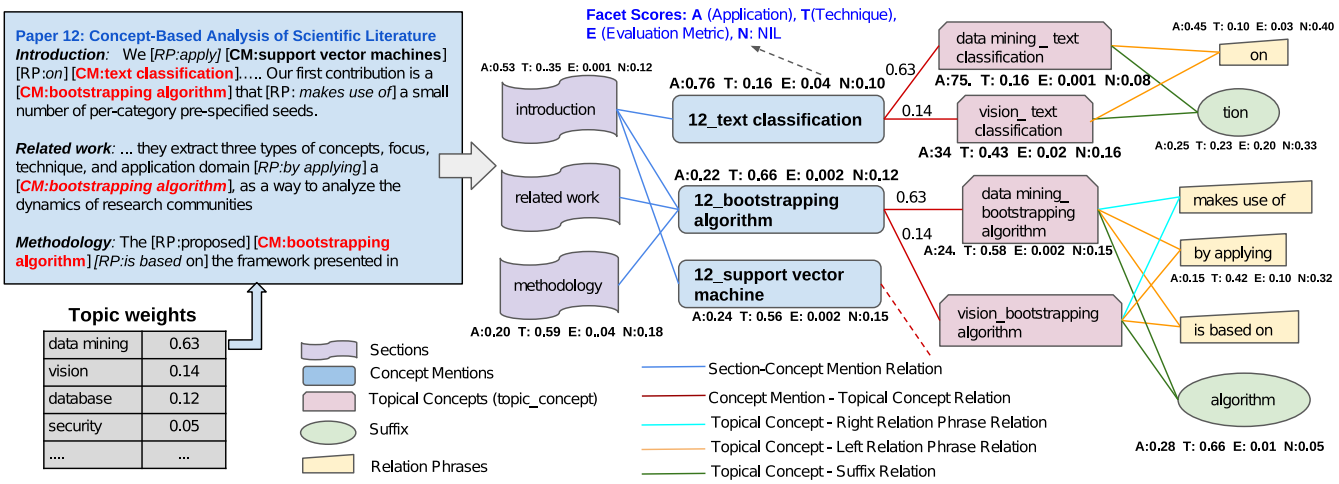


Figure 1: Overview of the FacetGist framework

(*e.g.*, "*document clustering*" as an application, "*SVM*" as a technique and "*DBLP bibliographic data*" as a dataset). A concept c can represent different facets in different documents. For example, while "*Decision Tree*" is used as a technique in documents related to data mining, it is used as an application in documents in machine learning field (see Fig. 3). Furthermore, even though a concept may represent different facets in different documents, it usually represents only one facet among its multiple mentions in a specific document. For each concept mention m, we use a $(l+1)$-dimensional binary indicator vector $\mathbf{y}_m \in \{0,1\}^{(l+1)}$ to indicate its facet category. $\mathbf{y}_{m,a} = 1$ if and only if the mention is of facet a. In particular, $\mathbf{y}_{m,l+1} = 1$ if the mention is Not-Of-Interest (NOI), i.e, it does not fall into any of facets of interests. By estimating $\mathbf{y}_m$, one can predict what facet the concept mention represents in its document as $\text{facet}(m) = \text{argmax}_{a \in \mathcal{A} \cup NOI}\, y_{m,a}$. Let $\mathcal{M} = \{m_1, \ldots, m_M\}$ denote the set of concept mentions in corpus $\mathcal{D}$. The facet indicators for $\mathcal{M}$ is denoted by $\mathbf{Y} \in \mathbb{R}^{M \times (l+1)}$.

Seed: Suppose the facets of a subset of mentions $\mathcal{M}_\mathcal{U} \subset \mathcal{M}$ (*i.e.*, *seed mentions*) can be confidently discovered using the word-based rules $\mathcal{U}$ that can automatically identify a few concepts for each facet (see Sec. 4). This work focuses on estimating the facets for the remaining mentions $\mathcal{M}_R = \mathcal{M} \setminus \mathcal{M}_\mathcal{U}$, consisting of two kinds of concepts: (1) concepts that are of the target facets in profile schema $\mathcal{A}$, and (2) concepts that are Not of Interest (NOI).

Problem Description. The input to Facet Extraction is a collection of documents $\mathcal{D}$, an *facet schema* $\mathcal{A} = \{a_1, \ldots, a_l\}$ where a_i is a facet of interest in documents, and a set of word-based rules $\mathcal{U}$ for seed generation. In our study, we assume all the mentions of a concept within a document are associated with a *single* facet or are not-of-interest, *i.e.*, $a \in \mathcal{A} \cup NOI$. We also assume $\mathcal{A}$ is given. It is beyond the scope of this document to generate $\mathcal{A}$. Formally, we define the problem of Facet Extraction as follows.

Definition 1 (Problem Definition). *Given a collection of documents $\mathcal{D}$, a facet schema $\mathcal{A}$ and a word-based rule set $\mathcal{U}$, our task is to: (1) extract candidate concept mentions $\mathcal{M}$ from $\mathcal{D}$; (2) generate seed mentions $\mathcal{M}_U$ with $\mathcal{U}$; and (3) for each document $d \in \mathcal{D}$, select a set of concept mentions $\mathcal{M}_{d,a}$ to represent each target facet $a \in \mathcal{A}$ of d, based on the estimated facet indicator vectors $\mathbf{Y}$.*

3. FACET EXTRACTION

At a high level, the proposed Facet Extraction framework consists of four major steps:

1. Performs a logical structure parsing and topic modeling on the document collection to obtain section structure and related topics for each document (Sec. 3.1).
2. Generates concept mentions by running a phrase segmentation algorithm and applying POS tag patterns and interestingness measure for filtering. Extracts relation phrases and suffix for each concept mention (Sec. 3.2)
3. Constructs a heterogeneous graph G to represent the available information between multiple extracted features in a unified form, which encodes our insights on modeling the key facets for each document (Sec. 3.3).
4. Estimates the facet indicator vectors for concept mentions across all documents collectively by solving the proposed joint optimization problem (Sec. 4).

3.1 Document Structure Extraction

Section Structure. Technical documents usually follow a logical structure to organize their content (*e.g.*, using sections, subsections, etc.). The sections within a document differ in their focus from each other. Intuitively, the section title provides hint on the facets of the concepts mentioned within the section. For example, concepts mentioned in "methodology" section are more likely to have the concepts corresponding to the Technique facet as compared to the concepts found in the "architecture" section.

We apply a conditional random field-based parsing tool, ParsCit [6], to retain the section titles for each document, and further group these titles into 9 major section categories (see Table 2) based on the labels predicted by ParsCit and regular expression rules. For example, to identify sections belonging to the *Introduction* category, we look for the headings that match [intro∗] or [overview∗]. These 9 categories represent more than 85% of the section titles found in our experimental datasets. All low frequency sections which do not fall into any of these sections are grouped under the *Miscellaneous* category.

Topic Structure. A collection of documents can be organized in terms of their topics (*i.e.*, fields of study such as *Information Retrieval*, *Machine Learning*, and *Computer Vision*). Topic modeling [5] derives such hidden theme structures for

abstract, introduction, methodology, system architecture, evaulation, data, example, conclusion, miscellaneous

Table 2: List of section categories used in this work for scientific documents

a given corpus. Each document can be categorized into multiple topics where each topic is represented by a multinomial distribution over all the unique words in the document collection. Intuitively, if two documents share similar topics (*e.g.*, both are about "product recommendation"), then the same concepts within the two documents tend to have the same facet (*e.g.*, "matrix factorization" mentioned in both documents represent the Technique facet of the documents).

Specifically, we apply latent Dirichlet allocation (LDA) [5] on the document collection $\mathcal{D}$ to generate a topic distribution $\boldsymbol{\theta}^i \in \mathbb{R}^K$ (*i.e.*, a multinomial distribution over the K topics with $\sum_{t=1}^{K} \theta_t^i = 1$) for each document $d_i \in \mathcal{D}$[1]. The number of topics K is decided based on the nature of the technical corpus (see Figure 4(d)). In doing so, we derive the weighted associations between documents and topics. As we will see in subsequent sections, the topics of the document where the concept is mentioned help in infering the facet of the concept.

3.2 Candidate Generation

Concept Mention. To ensure the generation of cohesive, informative, and salient concept mentions $\mathcal{M}$ for each document, we introduce a data-driven approach by incorporating both corpus-level statistics and local syntactic patterns. We first use a distantly-supervised phrase segmentation algorithm SegPhrase [17] to partition the text into non-overlapping segments to generate candidate concept mentions. Then we adopt the part-of-speech (POS) tag patterns (Table 4) and an *interestingness* measure [3] to guide the filtering of false concept mentions.

Given a word sequence, the result of phrase segmentation is a sequence of multi-word phrases or single words, each representing a cohesive content unit. Phrase segmentation represents each document as a *bag of phrases*, but only a few of these phrases make the representative concepts for the document. To select only high quality concept mentions, we filter out the phrases which do not have desirable POS tag patterns (*e.g.*, consecutive nouns). Further, we use an interestingness based metric to further remove the phrases which are not significant. The intuition behind interestingness is simple [3]: a phrase is more interesting to the document if it occurs frequently in the current document while relatively infrequently in the entire corpus. Let $\mathcal{P}_d$ denote the set of phrases which satisfy the POS patterns in Table 4 from the segmented document d, $n(p, d)$ denote the frequency of p in d, and $n(p, \mathcal{D})$ denote the document frequency of p in $\mathcal{D}$. The interestingness measure $I(\cdot)$ of p in $d \in \mathcal{D}$ is defined as follows [3].

$$I_{\mathcal{D}}(p, d) = \Big(0.5 + \frac{0.5 \times n(p, d)}{\max_{t \in \mathcal{P}_d} n(t, d)}\Big)^2 \cdot \log\Big(\frac{|\mathcal{D}|}{n(p, \mathcal{D})}\Big), \quad (1)$$

which is the product of the square of normalized term frequency and the inverse document frequency.

Table 3 depicts the effectiveness of our concept extraction method as compared to the noun-phrase chunking based concept extraction methods used by existing approaches [31]. Moreover, unlike the noun phrase extractor, our concept extractor treats two related words such as database systems and database management systems as the same concept.

[1]We use the LDA implementation in MALLET toolkit [18].

	DBLP			ACL		
Method	P	R	F1	P	R	F1
NP Chunker [22]	2.5	74.3	4.8	3.9	71.2	7.3
SegPhrase [17]	2.1	95.8	4.1	2.9	97.1	5.63
Ours	**26.1**	**90.5**	**40.5**	**27.7**	**87.75**	**42.1**

Table 3: Comparison with other concept extraction methods in terms of precision (P), recall (R) and F1 score.

Phrase Type	**POS tag patterns**
Concept Mention	N*, N*J, J
Relation Phrase	P, VP, VW*P

Table 4: POS patterns for filterng entiy mentions and relation phrases. V:Verb, N: Noun, P: Proposition, J=Adjective, W=N|V|P|J|Adverb

Note that a high recall in this stage is one of the key factors behind our overall improved performance. On further filtering out the noisy concept mentions using the interestingness threshold, we gain significant improvements in precision with a marginal decrease in recall.

Relation Phrase. A *relation phrase* is a phrase that denotes a unary or binary relation in a sentence [7]. Once we have identified a set of high quality concept mentions, we further identify the relation phrases in their left and right using the POS patterns depicted in Table 4. Extracting textual relations from documents has been previously studied [7] and applied to entity typing [19, 15]. We leverage the rich semantics embedded in relation phrases to provide facet cues for their concept argument. Specifically, we define the *facet signature* of a relation phrase p as two indicator vectors $\mathbf{p}_L, \mathbf{p}_R \in \mathbb{R}^{(l+1)}$. These vectors measure how likely the left/right concept arguments of p belong to different facets ($\mathcal{A}$ or NOI). A large positive value on $p_{L,a}$ ($p_{R,a}$) indicates that the left/right argument of p is likely of facet a.

Concept Suffix. We consider two kinds of patterns for concept suffix extraction: (1) common suffixes in English, and (2) frequent suffix unigrams (i.e last words in a concept phrases) extracted from the corpus. First, we use the 30 most common suffix patterns in English [2] which account for 93% of the candidate concept mentions in our experiments. These 30 suffixes are further grouped into 20 suffix groups based on their contextual usage. For example, we group together suffices like "*ition*", "*ation*", "*tion*", "*ion*" as these suffixes are generally appended to words for denoting some process or action. Similarly, we group together "*ible*" and "*able*". Second, we extract most frequent suffix unigrams from the candidate concept mentions $\mathcal{M}$ such as "algorithm", "model". As explained in the subsequent section, we link concepts that share these two suffix patterns in our constructed graph for facet propagation.

3.3 Construction of Heterogeneous Graphs

While most existing graph-based entity typing methods [15, 27, 24] rely on only relation phrases to infer entity types, we jointly model, using a heterogeneous graph, both the unstructured textual signals (*i.e.*, relation phrases, suffices) and the structured signals (*i.e.*, topics, sections) that are common to technical documents to infer the facets for all the concept mentions in the graph *collectively*. The basic idea for constructing the graph is that: *the more likely the two objects share the same facet, the larger the weight should be associated between their connecting edges*. We, now, formally define objects that make the graph and their relationships:

[2]http://www.darke.k12.oh.us/curriculum/la/suffixes.pdf

Objects: We use $\mathcal{R} = \{r_1, \ldots, r_p\}$ to denote p unique relations, $\mathcal{S} = \{s_1, \ldots, s_{|\mathcal{S}|}\}$ to $|\mathcal{S}|$ unique sections, and $\mathcal{X} = \{x_1, \ldots, x_{|\mathcal{X}|}\}$ to $|\mathcal{X}|$ unique suffixes extracted from the corpus $\mathcal{D}$. As the facet of a concept tends to differ across different topics, we introduce **topical concept** $z = t\_c$ to denote the role of a concept c within a topic t. A topical concept has more explicit facet than a concept. In other words, while a concept may change facet across documents, a concept across documents belonging to the same topic tends to have the same facet. For example, a topical concept "data mining_text classification" is most likely to be of facet Application for all documents. As we will see later in this section, we link each concept within a document to its all possible topical concepts with the weight of the link set to the weight of the topic in the document (see Figure 1). Thus, all documents that have a high weight for the topic "data mining" will tend to categorize the concept "text classification" as Application. We use $\mathcal{Z} = \{z_1, ..., z_{|\mathcal{Z}|}\}$ to denote the $|\mathcal{Z}|$ unique topical concepts extracted from the corpus. Further, we use an $(l+1)$-dimensional facet indicator vector $\mathbf{z} \in \mathbb{R}^{(l+1)}$ to measure how likely an object is subject to the l different facets in $\mathcal{A}$ or NOI.

Relationships: The constructed heterogeneous graph G has four types of links: *mention-concept link*, which represents the mapping between concept mention and topical concept, *mention-section link*, which models the document-level co-occurrences between concept mention and section, *concept-relation phrase link*, which captures corpus-level co-occurrences between topical concept and relation phrase, and *concept-suffix link*, which models the mapping between topical concept and its suffix.

Overall, we model the four kinds of relationships between concept mentions, topical concepts, relation phrases, sections and suffixes by the following hypothesis, and construct four subgraphs:

HYPOTHESIS 1 (GRAPH SMOOTHNESS). *If two objects in the graph are linked with a large weight, they tend to share the same facet (i.e., have similar facet indicator vectors).*

Mention-topical concept subgraph $G_{\mathcal{M},\mathcal{Z}}$: Intuitively, the topic distribution of the document, in which a concept mention occurs, provides cues on the association between the concept mention and its related topical concepts. If a concept mention m occurs in a document which is likely subject to topic t, then it tends to share the same facet indicator as the topical concept $z = t\_c$. For example, in Figure 1, there should be a strong association between the mention "*12_text classification*" and the topical concept "*data mining_text classification*" since "*data mining*" is a prevalent topic in document 12. Formally, the graph is represented using a bi-adjacency matrix $\mathbf{W}_{\mathcal{Z}} \in \mathbb{R}^{M \times |\mathcal{Z}|}$, where $W_{\mathcal{Z},ij} = \theta_t^d$. We denote the facet indicators for $\mathcal{Z}$ by matrix $\mathbf{Z} \in \mathbb{R}^{|\mathcal{Z}| \times (l+1)}$.

Mention-section subgraph $G_{\mathcal{M},\mathcal{S}}$: If a concept occurs relatively frequently in a specific section of a document, the concept mention should share similar facet indicator with the section's facet indicator (and vice versa). In Figure 1, since "*12_bootstrapping algorithm*" occurs more frequently in "*methodology*" (which in turn mostly contains concepts of facet Technique) than in other sections, then it is more likely referred as Technique in the document. Formally, the graph is represented using a bi-adjacency matrix $\mathbf{W}_{\mathcal{S}} \in \mathbb{R}^{M \times |\mathcal{S}|}$. Edge weight $\mathbf{W}_{\mathcal{S},ij} = n_d(m_i, s_j)/n_d(m_i)$ if m_i occurs in s_j, where we define $n_d(m_i, s_j)$ as frequency of m_i in s_j, and $n_d(m_i)$ as the frequency of m_i in d. We denote the facet indicators for $\mathcal{S}$ by $\mathbf{S} \in \mathbb{R}^{|\mathcal{S}| \times (l+1)}$.

Topical concept-relation phrase subgraph $G_{\mathcal{Z},\mathcal{P}}$: Between topical concepts and relation phrases, we exploit their co-occurrences aggregated across all documents in $\mathcal{D}$. If a concept c often appears as the left (right) argument of relation phrase p in documents with topic t, then facet indicator of z (where $z = t\_c$) tends to be similar to the corresponding facet indicator in p's type signature. We denote the facet indicators for left/right concept arguments of all relation phrases $\mathcal{P}$ by $\mathbf{P}_L, \mathbf{P}_R \in \mathbb{R}^{p \times (l+1)}$, respectively. In Figure 1, facet indicator of "*data mining_bootstrapping algorithm*" should be similar to $\mathbf{P}_R$ of the relation phrase "*makes use of*" as they co-occur frequently in the corpus. Formally, the graph is represented using two bi-adjacency matrices $\mathbf{W}_L$, $\mathbf{W}_R \in \mathbb{R}^{M \times p}$ to represent the co-occurrences between relation phrases and their left and right topical concept arguments, respectively. We define $W_{L,ij} = \sum_d \theta_t^d$ ($W_{R,ij} = \sum_d \theta_t^d$) if concept c of $z_i = t\_c$ occurs as the *closest* concept mention on the left (right) of relation phrase r_j in any document d in the corpus.

Topical concept-suffix subgraph $G_{\mathcal{Z},\mathcal{X}}$: If a topical concept contains a suffix, then the concept tend to have similar facet indicator as that of the suffix, and vice versa (*e.g.*, in Figure 1, facet indicators of "*data mining_boostrapping algorithm*" and suffix "*algorithm*" tend to be similar). Formally the graph is represented by a bi-adjacency matrix $\mathbf{W}_{\mathcal{X}} \in \{0,1\}^{n \times |\mathcal{X}|}$ where $W_{\mathcal{X},ij} = 1$ if topical concept $z_i = t\_c$ contains suffix x_j (*i.e.*, x_j is the suffix of c). We denote the facet indicators for $\mathcal{X}$ by $\mathbf{X} \in \mathbb{R}^{|\mathcal{X}| \times (l+1)}$.

To avoid overly popular objects in the subgraphs, we further normalize the rows and columns of all the relationship matrices $\{\mathbf{W}_{\mathcal{Z}}, \mathbf{W}_{\mathcal{S}}, \mathbf{W}_L, \mathbf{W}_R, \mathbf{W}_X\}$. If $\mathbf{W} \in \mathbb{R}^{n_1 \times n_2}$ represents any of the relationship matrices, the normalization is done as follows:

$$\mathbf{S} = \mathbf{D}^{(1)^{-1/2}} \cdot \mathbf{W} \cdot \mathbf{D}^{(2)^{-1/2}},$$

where we define the diagonal degree matrix $\mathbf{D}^{(1)} \in \mathbb{R}^{n_1 \times n_1}$ as $D_{ii}^{(1)} = \sum_{j=1}^{n_2} W_{ij}$, and the degree matrix $\mathbf{D}^{(2)} \in \mathbb{R}^{n_2 \times n_2}$ as $D_{jj}^{(2)} = \sum_{i=1}^{n_1} W_{ij}$. For example, to compute normalized matrix $\mathbf{S}_{\mathcal{Z}}$, we have $\mathbf{S}_{\mathcal{Z}} = \mathbf{D}^{(\mathcal{M})^{-1/2}} \cdot \mathbf{W}_{\mathcal{Z}} \cdot \mathbf{D}^{(\mathcal{Z})^{-1/2}}$.

4. FACET ESTIMATION ON GRAPHS

In this section, we introduce the facet estimation on the constructed heterogeneous graph.

As a simple solution, one can estimate the facet indicator vector for each kind of object (*i.e.*, topical concept, suffix and section) in the graph *separately*, and use the estimated vectors to predict the facet for concept mention $m \in \mathcal{M}$. However, such a solution does not fully leverage the data redundancy. For example, a concept may suffer from sparse co-occurrences with existing relation phrases while its relationships with suffices and sections can complement the facet estimation. In our solution, we formalize the facet estimation as a joint optimization problem which enforces the facet propagation between different kinds of objects in the graph *jointly*, by following our proposed hypothesis.

4.1 Supervision from Seed Concepts

We first introduce how to instantiate the heterogeneous graph (*i.e.*, generate seed concept mentions) using a set of word-based rules $\mathcal{U}$ provided by a human. We identified a set of suffix unigrams along with their related relation phrases to identify the candidate list of seed concepts of each facet. Table 5 depicts the unigrams and relation phrases that are used in generating seed concept mentions. We then manually verified and selected a few (we used 1000 for our experi-

Type	Terms
Application	retrieval, system, recognition, extraction, detection, resolution, generation
Technique	method, model, technique, algorithm, approach
Evaluation Metric	metric, score
Relation Phrase	in, for, based on, of, by, used in, using, apply, extend, proposed, train

Table 5: Terms used for selecting candidate seeds

ments) high-quality concept mentions among the candidates for each facet as the seed set for FacetGist. The second step is optional and is basically used to further improve the purity of seed concepts. In our experiments results (Figure 4(c)) we demonstrate that that beyond a certain minimum threshold, seed size has no major impact on the results.

Formally, a suffix word and relation phrase-based rule $u = (w, r/l, a) \in \mathcal{U}$ means that if one observes word w appears as its suffix in a concept mention $m \in \mathcal{M}$ and has r (or l) as its right (or left) relation phrase, then mention m is of facet $a \in \mathcal{A}$. Let function $I_u(m) : \mathcal{M} \mapsto \{0, 1\}$ indicate whether a mention m satisfies rule u. For each concept mention $m \in \mathcal{M}$, we define a seed vector $\mathbf{y}_m^{(0)} \in \mathbb{R}^{l+1}$, where $y_{m,i}^{(0)} = 1$ if $\sum_{u \in \mathcal{U}_i} I_u(m) > 0$ and $\mathcal{U}_i = \{u = (w, a) | a = a_i\}$, and $y_{m,i}^{(0)} = 0$ otherwise. Here we assume that mentions which do not satisfy any rule will have all zeros in their seed vectors (*i.e.*, no prior knowledge on their facets). We model the seed information on concept mentions as follows.

$$\mathcal{H}(\mathbf{Y}, \mathbf{Y}^{(0)}) = \frac{1}{2} \sum_{i=1}^{M} \|\mathbf{y}_{m_i} - \mathbf{y}_{m_i}^{(0)}\|_2^2 = \frac{1}{2} \|\mathbf{Y} - \mathbf{Y}^{(0)}\|_F^2, \quad (2)$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. Minimizing this term enforces the estimated facet vector for concept mentions to be similar to the seed facet vector that encodes the prior knowledge on facets of concept mentions.

4.2 Modeling the Constructed Graph

With the constructed graph, a key component in formalizing the optimization problem is to model the proposed hypotheses in Sec. 3. The basic idea behind our proposed hypotheses is simple: Two objects are likely to share similar facet indicator vectors (*i.e.*, similar confidence score in terms of each facet) if and only if there exists a strong association between them (*i.e.*, large link weight connected between them). We leverage graph-based semi-supervised learning [34, 33] to model this idea. It enforces that two connected objects have similar facet vectors by preserving the intrinsic manifold structure among them (*i.e.*, graph consistency). Such a graph consistency term can be used to model each subgraph in the constructed heterogeneous graph. We further take the weighted combination of multiple graph consistency terms to jointly model the facet propagation over a heterogeneous graph (Sec. 4.3).

Specifically, suppose we have a bi-adjacency matrix $\mathbf{W} \in \mathbb{R}^{n_1 \times n_2}$ to represent a bipartite subgraph G_{12} between two kinds of objects $\mathcal{E}_1$ and $\mathcal{E}_2$. Let $\mathbf{Y}^{(1)} \in \mathbb{R}^{n_1 \times (l+1)}$ and $\mathbf{Y}^{(2)} \in \mathbb{R}^{n_2 \times (l+1)}$ be the facet indicator matrices for objects $\mathcal{E}_1$ and $\mathcal{E}_2$ respectively. We define the graph consistency term $\mathcal{L}_{12}$ for subgraph G_{12} as follows.

$$\mathcal{L}_{12}(\mathbf{Y}^{(1)}, \mathbf{Y}^{(2)}, \mathbf{W}) = \frac{1}{2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} W_{ij} \Big\| \frac{\mathbf{y}_i^{(1)}}{\sqrt{D_{ii}^{(1)}}} - \frac{\mathbf{y}_j^{(2)}}{\sqrt{D_{jj}^{(2)}}} \Big\|_2^2;$$
$$= \mathbf{Y} - \mathbf{Y}^{(1)T} \cdot \mathbf{S} \cdot \mathbf{Y}^{(2)}. \quad (3)$$

Here, $\mathbf{D}^{(1)} \in \mathbb{R}^{n_1 \times n_1}$ is the diagonal degree matrix for objects $\mathcal{E}_1$, defined as $D_{ii}^{(1)} = \sum_{j=1}^{n_2} W_{ij}$, and $\mathbf{D}^{(2)} \in \mathbb{R}^{n_2 \times n_2}$ is the degree matrix for objects $\mathcal{E}_2$ where $D_{jj}^{(2)} = \sum_{i=1}^{n_1} W_{ij}$. We use $\mathbf{S}$ to denote the normalized matrix of $\mathbf{W}$ where $\mathbf{S} = \mathbf{D}^{(1)-\frac{1}{2}} \cdot \mathbf{W} \cdot \mathbf{D}^{(2)-\frac{1}{2}}$.

By minimizing the term in Eq. (3), a larger link weight between two objects (*i.e.*, W_{ij}) will enforce the difference term $\|\cdot\|_2$ to be small, *i.e.*, two facet indicators to be similar to each other, which models the idea of graph consistency.

4.3 The Joint Optimization Problem

We now focus on formulating a joint optimization problem to unify different subgraphs (relations) in the constructed heterogeneous graph G. In the objective function, each subgraph is encoded into one graph consistency term (*i.e.*, Eq. (3)) to preserve the intrinsic manifold structure between the objects, *i.e.*, two linked objects tend to have similar confidence scores in their facets according to the strength of relationship between them. Different terms are then combined with the corresponding weights which trade-off the strength of signal between each subgraph, and they are further unified with the supervision term in Eq. (2) to form the objective function as follows.

$$\begin{aligned} \mathcal{O} &= \alpha \mathcal{H}(\mathbf{Y}, \mathbf{Y}^{(0)}) + \lambda_{\mathcal{MZ}} \cdot \mathcal{L}_{\mathcal{MZ}}\big(\mathbf{Y}, \mathbf{Z}, \mathbf{W}_{\mathcal{Z}}\big) \quad (4) \\ &+ \lambda_{\mathcal{ZS}} \cdot \mathcal{L}_{\mathcal{MS}}\big(\mathbf{Y}, \mathbf{S}, \mathbf{W}_{\mathcal{S}}\big) + \lambda_{\mathcal{ZP}} \cdot \mathcal{L}_{\mathcal{ZP}_L}\big(\mathbf{Z}, \mathbf{P}_L, \mathbf{W}_L\big) \\ &+ \lambda_{\mathcal{ZP}} \cdot \mathcal{L}_{\mathcal{ZP}_R}\big(\mathbf{Z}, \mathbf{P}_R, \mathbf{W}_R\big) + \lambda_{\mathcal{ZX}} \cdot \mathcal{L}_{\mathcal{ZX}}\big(\mathbf{Z}, \mathbf{X}, \mathbf{W}_{\mathcal{X}}\big) \\ &+ \beta\big(\|\mathbf{Y}\|_F^2 + \|\mathbf{Z}\|_F^2 + \|\mathbf{S}\|_F^2 + \|\mathbf{P}_L\|_F^2 + \|\mathbf{P}_R\|_F^2 + \|\mathbf{X}\|_F^2\big). \end{aligned}$$

Here, a set of tuning parameters $0 < \{\lambda_{\mathcal{MZ}}, \lambda_{\mathcal{ZP}}, \lambda_{\mathcal{ZX}}, \lambda_{\mathcal{MS}}\} < 1$ are used to control the strength of signals in different subgraphs in the objective function. Furthermore, we use a tuning parameter $0 < \alpha < 1$ to trade-off between the relations in heterogeneous graph and supervision from seed concept mentions. To avoid trivial solutions, we add L_2 regularization for variables $\{\mathbf{Y}, \mathbf{Z}, \mathbf{S}, \mathbf{P}_L, \mathbf{P}_R, \mathbf{X}\}$.

In Eq. (4), the first term propagates the facet information from seeded concept mentions to the remaining ones. The second term models relations between topical concepts and sections. The third and fourth terms enforce Hypothesis 1 for topical concepts and relation phrases. The fifth term models suffix information for topical concepts.

To derive the exact type of each candidate concept mention, we impose the 0-1 integer constraint $\mathbf{Y} \in \{0, 1\}^{M \times T}$ and $\mathbf{Y1} = \mathbf{1}$. With the definition of $\mathcal{O}$, we define the joint optimization problem as follows.

$$\min_{\substack{\mathbf{Y}, \mathbf{Z}, \mathbf{S} \\ \mathbf{P}_L, \mathbf{P}_R, \mathbf{X}}} \mathcal{O} \quad \text{s.t. } \mathbf{Y} \in \{0, 1\}^{M \times (l+1)}, \quad \mathbf{Y1} = \mathbf{1}. \quad (5)$$

The facet indicator matrix $\mathbf{Y}$ can be estimated by solving the joint optimization problem in Eq. (5). We can then predict the more likely facet for each concept mention based on $\text{facet}(m) = \text{argmax}_{a \in \mathcal{A} \cup NOI} \; y_{m,a}$ for $m \in \mathcal{M}$.

4.4 An Iterative Algorithm

The optimization problem in Eq. (5) is mixed-integer programming and is NP-hard to solve. Instead of solving it directly, we propose a two-step approximate algorithm. The algorithm first solves the real-valued relaxation of Eq. (5), *i.e.*, $\mathbf{Y} \in \mathbb{R}^{M \times (l+1)}$; then it imposes the binary constraints

Algorithm 1: Facet Extraction

Input: heterogeneous graph G, tuning parameters $\{\alpha, \lambda_{\mathcal{MZ}}, \lambda_{\mathcal{ZP}}, \lambda_{\mathcal{ZX}}, \lambda_{\mathcal{MS}}, \beta\}$
Output: Top concepts for each facet a of paper $d \in \mathcal{D}$

Initialize: $\{\mathbf{Y}^0\}$ by word/relation-based seed rules; $\{\mathbf{Y}, \mathbf{Z}, \mathbf{P}_L, \mathbf{P}_R, \mathbf{S}, \mathbf{X}\}$ as zero matrices
repeat
  Update $\mathbf{Z}$ using Eq. (7)
  Update $\{\mathbf{P}_L, \mathbf{P}_R\}$ using Eq. (8)
  Update $\mathbf{Y}$ following Eq. (9)
  Update $\mathbf{X}$ and $\mathbf{S}$ using Eq. (10)
until *the objective* $\mathcal{O}$ *in Eq.* (4) *converges*
return *estimated* $\mathbf{Y}$
Rank concepts under each facet a for each paper d according to concept's interestingness score;

back to predict the exact facet of each concept mention $m_i \in \mathcal{M}$ by selecting the element with the highest confidence score, *i.e.*, $\text{facet}(m) = \text{argmax}_{a \in \mathcal{A} \cup NOI}\ y_{m,a}$.

Closed-formed solution. With all variables in Eq. (4) being real-valued, we can rewrite the objective into a convex objective as follows.

$$\mathcal{O}_{relaxed} = \mathbf{F}^T\mathbf{M}\mathbf{F} + \mathbf{F}^T\mathbf{\Omega}\mathbf{F} - 2 \cdot \mathbf{F}^T\mathbf{\Omega}\mathbf{I}_0, \tag{6}$$

where $\mathbf{F} = [\mathbf{Y}^T, \mathbf{Z}^T, \mathbf{P}_L^T, \mathbf{P}_R^T, \mathbf{X}^T, \mathbf{S}^T]^T$ is the augmented facet indicator matrix for all objects, $\mathbf{M}$ is a symmetric matrix based on the set of bi-adjacency matrices $\{\mathbf{W}_{\mathcal{Z}}, \mathbf{W}_L, \mathbf{W}_R, \mathbf{W}_{\mathcal{X}}, \mathbf{W}_{\mathcal{S}}\}$, $\mathbf{\Omega}$ is a diagonal matrix based on α, and $\mathbf{I}_0 = [\mathbf{Y}^{(0)T}, \mathbf{0}, \mathbf{0}, \mathbf{0}]^T$ is the augmented seed matrix.

We can prove that $\mathbf{M}$ is positive semi-definite (and thus invertible) by referring to Eqs. (4) and (3). By taking the derivative with respect to $\mathbf{F}$ and setting it to zero, we can derive the closed-form solution for relaxed optimization problem as $\mathbf{F}^* = (\mathbf{M}+\mathbf{\Omega})^{-1}\mathbf{\Omega} \cdot \mathbf{I}_0$. The convex optimization problem has the closed-form solution as its global minimum.

Iterative update formula. However, directly computing the closed-form solution requires us to inverse the matrix $\mathbf{M}+\mathbf{\Omega}$, an intractable problem when handling a large heterogeneous graph. Instead, we provide an efficient iterative algorithm that generates the same globally minimal solution.

With facet to confidence scores $\mathbf{Y}$ for concept mentions initialized by seed mentions as in $\mathbf{Y}^{(0)}$, for each variable in $\{\mathbf{Y}, \mathbf{Z}, \mathbf{S}, \mathbf{P}_L, \mathbf{P}_R, \mathbf{X}\}$, we iteratively update its facet indicator matrix by fixing the values of other variables (as their previous values). The update formula can be derived by taking derivative of $\mathcal{O}$ in Eq. (4) with respect to each variable.

$$\mathbf{Z} = \frac{1}{\beta}\big[\lambda_{\mathcal{ZP}} \cdot \mathbf{S}_L\mathbf{P}_L + \lambda_{\mathcal{ZP}} \cdot \mathbf{S}_R^T\mathbf{P}_R + \lambda_{\mathcal{CX}} \cdot \mathbf{S}_{\mathcal{X}}\mathbf{X}\big]; \tag{7}$$

$$\mathbf{P}_L = \frac{1}{\beta}\mathbf{S}_L^T\mathbf{Z}; \quad \mathbf{P}_R = \frac{1}{\beta}\mathbf{S}_R^T\mathbf{Z}; \tag{8}$$

$$\mathbf{Y} = \frac{1}{2\gamma}\big[\lambda_{\mathcal{MZ}} \cdot \mathbf{S}_{\mathcal{Z}}\mathbf{Z} + \lambda_{\mathcal{MS}} \cdot \mathbf{S}_{\mathcal{S}}\mathbf{S} + \alpha \cdot \mathbf{Y}^{(0)}\big]; \tag{9}$$

$$\mathbf{X} = \frac{1}{\beta}\mathbf{S}_{\mathcal{X}}^T\mathbf{Z}; \quad \mathbf{S} = \frac{1}{\beta}\mathbf{S}_{\mathcal{S}}^T\mathbf{Y}; \tag{10}$$

where $\gamma = \alpha + \lambda_{\mathcal{MZ}} + 2\lambda_{\mathcal{CP}} + \lambda_{\mathcal{CX}} + \lambda_{\mathcal{MS}}$.

Algorithm 1 summarizes the iterative algorithm. In the iterative solution, each object iteratively spreads its score to its neighbors in the heterogeneous graph following hypotheses in Section 3.3 until a global stable state is achieved. Following an analysis similar to [33], one can prove that our algorithm converges to the closed-form solution in Eq. (6).

5. EXPERIMENTS

Experimental Settings: To verify the effectiveness of our proposed model, we conducted experiments on two real data sets: **DBLP** [3] and **ACL** [4]. As explained in Section 3.1, our heterogeneous graph model is built out of the following entities: *sections, topics, concept mentions* and *relation phrases* which are extracted from the corpora. We tried different numbers of topics K for each dataset. While keeping other factors of propagation fixed, we found $K = 15$ in DBLP and $K = 9$ in ACL gives in the best F1 scores (Figure 4(d)). We used SegPhrase [17], and applied POS tag patterns and interesting based filtering to extract concept menitons. Using a validation set of 25 documents, we set maximal pattern length, minimum support and significance threshold in SegPhrase to 8, 8 and 2 in ACL and 5, 14 and 2 in DBLP. We set interestingness threshold to 0.08 in ACL and 0.06 in DBLP. Performance results of our concept extractor are provided in Section 3.2. We set $(\alpha, \lambda_{\mathcal{MZ}}, \lambda_{\mathcal{ZP}}, \lambda_{\mathcal{ZX}}, \lambda_{\mathcal{MS}}, \beta)$ to $(0.5, 0.1, 0.1, 0.1, 0.1, 1)$ for both the datasets. We selected a subset of diverse documents (150 for DBLP and 75 for ACL) and took help from two human subject matter experts to annotate them with concepts belonging to three target facets: Application, Technique, and Evaluation Metric. The inter-rater agreements (kappa-value) between the experts, were 93.2% on DBLP and 89.0% on ACL. Concept mentions with conflicting facets were further discussed and resolved together by the two experts. Table 5 contains the unigrams and relation phrases that were used to generate the candidate list for seeds, and then from the candidate list, 1000 high-quality seeds across the three facets were manually selected. Table 6 provides more details about the datasets, graph entities, and the gold truth.

Data Sets	DBLP	ACL
#Documents	51897	11203
#Concept Mentions after POS Filtering	1.34m	340k
#Concept Mention after Interestingess Filtering	310k	128k
#Relation Phrases	88k	34k
#Sections	232k	45k
#Topics	15	9
#Gold Truth documents	150	75
#Gold Truth Technique	994	580
#Gold Truth Application	570	258
#Gold Truth Evaluation Metric	78	54

Table 6: Dataset, Ground Truth and Graph Statistics

	DBLP			ACL		
Method	P	R	F1	P	R	F1
Baseline	44.8	46.2	45.4	40.8	44.4	42.5
FacetGist	**73.5**	**74.1**	**73.7**	**67.1**	**72.3**	**69.6**

Table 7: Precision (P), Recall (R) and F1 scores: FacetGist vs Baseline

Method	DBLP								
	Application			Technique			Evaluation Metric		
	P	R	F1	P	R	F1	P	R	F1
Baseline	35.3	42.9	38.7	43.1	45.2	44.1	73.2	76.5	74.8
FacetGist: S+C	56.3	27.5	36.9	50.3	29.6	37.2	55.2	40.3	46.5
FacetGist: Se+C	72.7	25.6	37.8	76.3	35.6	48.5	62.3	74.2	67.7
FacetGist: R+C	60.2	51.2	55.3	63.5	64.4	63.9	59.2	67.8	63.2
FacetGist: R+C+To	63.6	58.2	60.7	67.4	68.9	68.1	60.3	69.4	64.5
FacetGist: all	68.3	69.3	68.7	76.9	78.2	77.5	75.2	77.9	76.5

Table 9: DBLP: Precision, Recall and F1 scores comparison of FacetGist with baseline on individual facets while varying entities; S:Suffix, C: Concept Mentions, Se: Sections, R: Relation Phrases, To: Topics

Baseline: The bootstrapping based concept extraction approach proposed by Tsai et al. [31] is currently the state-of-the-art technique for concept extraction in the scientific

[3]DBLP dataset:https://datahub.io/dataset/dblp
[4]ACL dataset: http://acl-arc.comp.nus.edu.sg

Topic	Top 10 Application	Top 10 Technique
Databse	relational database (408), query processing (390), database system (368), query evaluation (315), query execution (289), query language (276), xml documents (189), query optimization (172), object oriented databases (165), concurrency control (164)	data modeling (243), entity recognition (228), association rules (170), relational model (164), hash join (145), graph search (120), heirarchical model (116), parallel processing (110), k-nearest neighbour (105), pattern matching (84)
NLP/ Information Extraction	information extraction (442), information retrievel (275), natural language (264), pos tagging (257), machine translation (190), extraction task (183), relation extraction (172), speech recognition (148), pronoun resolution (98), translation systems (84)	language model (306), translation model (220), gram model (172), feature selection (169), proabilistic model (166), translation model (162), machine learning (161), topic model (154), prediction model (143)
Machine Learning	machine learning (270), model selection (232), learning problem (209), feature selection (204), information retrieval (152), learning process (148), clustering (137), density estimation (129), classification problem (127), bayesian inference (115)	generative model (395), objective function (342), mixture model (307), probabilistic model (292), em algorithm (188), kernel function (172), topic model (157), linear combination (144), language model (128), gradient descent (122)
Data mining	classification (378), object recognition (189), clustering process (168), image classification (162), online learning (128), transfer learning (115), face recognition (112), social network (109), anomaly detection (96), optimization problem (94)	feature selection (322), machine learning (247), support vector machines (186), decision tree (181), learning algorithm (158), neural networks (134), kernel methods (120), generative model (118), dimensionality reduction (98), feature learning (98)

Table 8: Top 10 concepts in Application and Technique for selected topics

Method	ACL								
	Application			Technique			Evaluation Metric		
	P	R	F1	P	R	F1	P	R	F1
Baseline	39.6	42.9	41.1	43.1	45.2	44.1	68.5	70.6	69.5
FacetGist: S+C	58.1	24.7	34.6	66.9	25.3	36.7	28.4	37.5	32.3
FacetGist: Se+C	74.1	27.9	40.5	73.4	30.8	43.3	43.3	49.8	46.3
FacetGist: R+C	68.8	42.1	52.2	52.8	54.8	53.7	58.1	55.3	56.6
FacetGist: R+C+To	72.5	49.5	58.8	59	64.3	61.5	61.2	65.3	63.1
FacetGist: all	63.9	71.2	67.3	68.9	73.9	71.3	75.4	78.2	76.7

Table 10: ACL: Precision, Recall and F1 scores comparison of FacetGist with baseline on individual facets while varying entities

literature. Similar to our method, the bootstrapping algorithm uses a small number of pre-specified seeds of each target facet. From the concept mentions (noun phrases) that contain the unigrams mention in the seed list (Table 5 – we use the same terms to select the candidate seeds in our method), the baseline extracts features such as unigrams, bigrams, left unigram, right unigram, left bigram, right bigram, closest verb, and the capitalization for each facet. These extracted features are used to annotate more concept mentions, which in turn are used for extracting additional features. This step is repeated until no new features are added. The final set of features are used to label the facet of concept mentions in the test set. Using a development set of 25 documents in both datasets, we set parameters (k, n, t) (the symbols have same defintions as in [31]) to $(2000, 200, 2)$.

Evaluation Metrics: We use F1 score computed from Precision(P) and Recall(R) to evaluate the concept extraction and facet identification performance for each of the facet. We denote the predicted concept mentions as J and the ground truth annotated mentions in the evaluation set as A. Precision is calculated as $\frac{\#(J \cap A)}{\#J}$ and Recall is calculated as $\frac{\#(J \cap A)}{\#A}$.

5.1 Performance results

Facet Identification: As depicted in Table 7, FacetGist results in a considerable improvement of 25-30% in the F1 score over the baseline. This is due to multiple factors. First, FacetGist indentifies many more concepts in the concept extraction phase than the noun-phrase based concept extractor used by the baseline (see recall in Table 3). Second, multiple local (relation phrase and suffix) and global signals (sections, topics) reduce the sparsity of the constructed graph by adding more links between the concept mentions. Moreover, the signals together better capture the context in which a concept is used than just the local sentence level features used in the baseline. Fianlly, global signals help in facet disambiguation when the concepts change roles across documents, while the baseline maps each concept to only one facet throughout the corpus.

Table 9 and 10 depict the contribution of individual factors in facet propagation. Suffixes or section-based propagations result in moderate precision but lower recall since they are insufficient in completely capturing the context of a concept mention. Relation phrases play the major role (depicting an improvement of about 25-45% in F1 score over only suffix or only section based propagation) in facet identification as concepts belonging to similar facets generally have similar phrases on their left and right. Moreover, relation phrases create many more links among the concept mentions, and thus help in reducing the overall sparsity of the graph for a better facet propagation. However, since relation phrases cannot capture the global context of concept. To add global signals, we combine topical-concepts with relation phrases which result in a substantial jump (10%) in precision, mainly because of the facet disambiguation. Finally, we do a joint propagation of suffixes, sections, relation phrase and topical-concepts as per the algorithm outlined in Section 4.4 which results in an improved F1 score over individual subgraph propagation by about 15-25%. Similar results are observed across all facets in both the datasets.

Paper Title	Top 10 concepts along with their facets
Concept-based analysis of scientific literature [31]	scientific literature (APP), unsupervised learning (TECH), trend analysis(APP), bootstrapping algorithm (TECH), text classification (APP), F1 score (EVAL), concept clustering (TECH), precision (EVAL), concept extraction (APP)
ClusCite: effective citation recommendation by information network-based clustering [25]	citation recommendation (APP),clustering (TECH), recall (EVAL), cluscite algorithm (TECH), heterogeneous bibliographic network (TECH),precision (EVAL), behavioral pattern (APP), mean reciprocal rank (EVAL), authority propagation (TECH)
Identifying relations for open information extraction [7]	information extraction (APP), reverb (TECH), textrunner (TECH), informative extraction (APP), incoherent extraction (APP), extraction algorithm (TECH), textrunner r (TECH), precision (EVAL), recall (EVAL), relation extraction (APP), AUC (EVAL)

Table 11: Extracted concepts along with their facets on three example papers

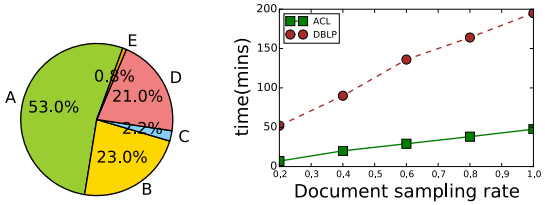


Figure 2: Left:Time taken by each component on DBLP (similar results on ACL): A) Parsing and Section Extraction, B) Topic Extraction, C) Graph Construction D) Candidate Extraction (DBLP), E) Type Propagation; Right: Time Vs Corpus size (by sampling at various rates)

Effect of corpus size: Figures 4(a) and 4(b) depict the performance of our method when varying the size of the corpus. We vary the size of the corpus by varying the sampling rate of documents from the overall corpus. Since ACL is a relatively smaller corpus (11k docs), a low sampling rate results in a sparse subgraph causing poor facet propagation. As the sampling ratio is increased, there is a significant improvement in F1 score. After a certain point, when the number of documents is enough in the corpus, the subgraphs are more connected and further increasing of the corpus size does not affect the F1 score. For DBLP, the F1 score is high from the very beginning as there are enough number of doc-

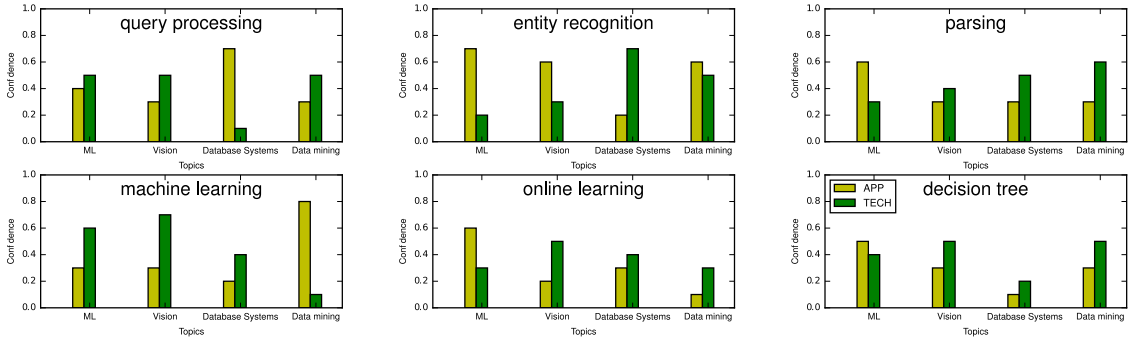
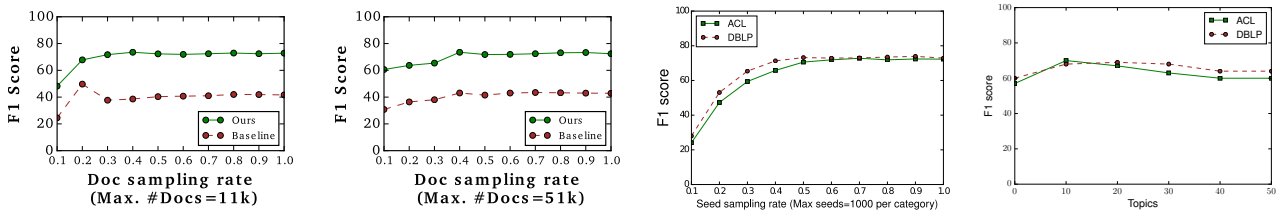


Figure 3: Variation of facets across Topics for selected concepts



(a) Varying #Docs (ACL) (b) Varying #Docs (DBLP) (c) Varying Seed Size (d) Varying Topic Size

Figure 4: Performance on varying the doc size, seed size and topic size

uments even when the sampling rate is low. A similar trend is observed for the baseline.

Effect of the size of seed mention set: Seed mentions are used for instantiating our heterogeneous graph. To see the effect of seed mentions, we vary the size of seeds by sampling at different rates from a maximum of 1000 seeds and observe the change in F1 scores. Figure 4(c) shows that on increasing the size of seed mentions, F1 score improves drastically up to a certain sampling ratio, after which there is no significant improvement. Improvement in the initial stage is mainly due to the increase in recall.

Effect of the number of topics To see the effect of the number of topics on overall results, we varied the number of topics from 0 to 50. Figure 4(d) shows that F1 score is the highest for DBLP when the number of topics is between 10 and 20 while for ACL, it is the highest around 10. Both a too low or a too high number of topics affects the facet-disambiguation resulting in poor precision.

5.2 Case Studies

1. Example output on three papers Table 11, depicts top 10 concepts (ranked according to their interestingness score) with their facets in three documents from DBLP. These concepts capture the most important Technique, Application and Evaluation Metric in the documents.

2. Top Application and Technique in different topics Table 8 depicts the most common Application and Technique (ranked according to their frequency) for 4 selected topics in DBLP. The frequency of concepts for each topic are calculated as follows: for each document, we multiply the term-frequency of a concept with the topic weight to get the frequency of topical-concept per document. Then, for each topical-concept, we sum the frequency across documents to obtain the overall frequency. There are two interesting observations. First, the top concepts under each facet for different topics are among the key concepts in the sub-domain reflected by the topic. Second, the same concept changes role across different topics, demonstrating the effectiveness of our model in facet disambiguation.

3. Ambiguous facets across topics. In Figure 3, we show variations in facets across topics for six example concepts. we calculate the facet score for a topical concept as follows: we find the facet score for each topical-concept by summing the scores of topical-concepts scores across all papers for the facet type. Then for each topical concept, we normalize the scores across all facets. Figure 3 depicts the variation of facet scores of concepts such as query processing, online learning, parsing across different topics. For instance, it can be seen that machine learning concept has a high score of being an Application under machine learning topic, while it is used as a Technique under the data mining topic.

5.3 Test of Scalability

1. Time cost change w.r.t. corpus size. We performed our experiments on a machine with 20 cores of Intel(R) Xeon(R) CPU E5-2680 v2 @ 2.80GHz. Our framework is implemented in Python. As depcited in Figure 2, the total time for running our method varies linearly as compared to the size of the corpus.

2. Time cost ratio of different modules. Figure 2 depicts the ratio of time taken by different components of our framework. We observe that document parsing to identify sections takes the maximum amount of time followed by candidate extraction and topic modeling. facet propagation takes the least amount of time—up to 1% of the total time.

6. RELATED WORK

Our method is, in general, related to existing work on entity recognition, specifically weakly supervised entity classification techniques [14, 29]. Using a seed list of high-quality entities, weakly-supervised methods can extract more entities and relations of target types in a fast and cheap manner. These methods rely only on local textual clues, and try to recognize and type all entity mentions whereas we utilize sentence, document, and corpus level cues to select a subset of important keyphrases, and categorize and rank them for given facets.

ClusType [24], similar to our method, constructs a heterogeneous network-based model of entity mentions and relation phrases, and applies type propagation and relation phrase clustering in a mutually enhanced manner to predict concepts belonging to specific types. However, ClusType does not take advantage of contextual signals beyond the

reach of relation phrase clusterings, such as the document and corpus-level signals important to the technical domain.

Several bootstrapping-based methods [31, 10, 30, 28, 20] have been proposed for identifying important concepts in medical domain and scientific corpus. [31] is currently the state-of-the-art, and also the baseline and starting point for our work. Most of these work make use of local sentence level clues and assign a single facet to a concept throughout the corpus whereas our method exploits both local and global features, and can disambiguate facets for a concept across document. Moreover, we incorporate corpus-level statistics and document-level interestingness metric to filter out the unimportant concepts. We further rank the extracted concepts according to their interestingness within each facet.

There has been some work [16, 8] on aspect based sentiment analysis, where the goal is to identify sentiments (positive or negative) expressed for each aspect (e.g., battery life) of target entities (e.g., mobile), while our work identifies concepts for key aspects such as techniques, applications for each document instead of sentiments. Similarly, attribute mining [12] aims to group together multiple attributes or entities belonging to similar concept hierarchy such as company and country. We, on the other hand, identify concepts at a document level, and the same concept in our work can also change facets across documents.

7. CONCLUSION

In this paper, we define and formalize the novel problem of Facet Extraction in corpora of technical documents. We propose a weakly-supervised framework that integrates both local context signals (*e.g.*, relation phrases, concept suffix, etc.) with global structure signals (*e.g.*, paper sections, and topics) in a unified heterogeneous graph-based data model. We then formulate a joint optimization problem for estimating the facets of the candidate concepts, following the idea of graph-based label propagation. Our experiments on real-world datasets demonstrate the effectiveness of the proposed approach. In future work, we plan to further extend our model to incorporate more structure signals such as citations and co-author relationships. We also plan to apply concept and relation phrase clustering to merge semantically similar phrases. This would help in decreasing the overall sparsity of the graph for a better facet propagation.

Acknowledgements

Research was sponsored in part by the U.S. Army Research Lab. under Cooperative Agreement No. W911NF-09-2-0053 (NSCTA), National Science Foundation IIS-1320617, IIS-1354329, IIS 16-18481, IIS-1513407 and IIS-1633755, and HDTRA1-10-1-0120, and grant 1U54GM114838 awarded by NIGMS and 3U54EB020406-02S1 awarded by NIBIB through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov), the Faculty Research Award provided by Google, and a grant from the Siebel Energy Institute. The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies of the U.S. Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation hereon.

References

[1] D. Aumüller and E. Rahm. Affiliation analysis of database publications. *SIGMOD Record*, 40(1):26–31, 2011.

[2] J. Ayres, J. Flannick, J. Gehrke, and T. Yiu. Sequential pattern mining using a bitmap representation. In *SIGKDD*, 2002.

[3] S. Bedathur, K. Berberich, J. Dittrich, N. Mamoulis, and G. Weikum. Interesting-phrase mining for ad-hoc text analytics. *VLDB*, 3(1-2):1348–1357, 2010.

[4] S. Bethard and D. Jurafsky. Who should I cite: learning literature search models from citation behavior. In *CIKM*, 2010.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[6] I. G. Councill, C. L. Giles, and M.-y. Kan. Parscit: An open-source crf reference string parsing package. In *LREC*, 2008.

[7] A. Fader, S. Soderland, and O. Etzioni. Identifying relations for open information extraction. In *EMNLP*, 2011.

[8] R. Feldman. Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89, 2013.

[9] J. Gao, F. Liang, W. Fan, C. Wang, Y. Sun, and J. Han. On community outliers and their efficient detection in information networks. In *SIGKDD*, 2010.

[10] S. Gupta and C. D. Manning. Analyzing the dynamics of research by extracting key aspects of scientific papers. In *IJCNLP*, 2011.

[11] S. Gupta and C. D. Manning. Improved pattern learning for bootstrapped entity extraction. In *CONLL*, 2014.

[12] A. Halevy, N. Noy, S. Sarawagi, S. E. Whang, and X. Yu. Discovering structure in the universe of attribute names. In *Proc. 25th International Conference on World Wide Web (WWW)*, 2016.

[13] K. S. Hasan and V. Ng. Automatic keyphrase extraction: A survey of the state of the art. *ACL*, 2014.

[14] Z. Kozareva, K. Voevodski, and S.-H. Teng. Class label enhancement via related instances. In *EMNLP*, 2011.

[15] T. Lin, O. Etzioni, et al. No noun phrase left behind: detecting and typing unlinkable entities. In *EMNLP*, 2012.

[16] B. Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

[17] J. Liu, J. Shang, C. Wang, X. Ren, and J. Han. Mining quality phrases from massive text corpora. In *SIGMOD*, 2015.

[18] A. K. McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[19] N. Nakashole, T. Tylenda, and G. Weikum. Fine-grained semantic typing of emerging entities. In *ACL*, 2013.

[20] K. Nigam and R. Ghani. Analyzing the effectiveness and applicability of co-training. In *CIKM*, 2000.

[21] A. Parameswaran, H. Garcia-Molina, and A. Rajaraman. Towards the web of concepts: Extracting concepts from large datasets. *VLDB*, 3(1-2):566–577, 2010.

[22] V. Punyakanok and D. Roth. The use of classifiers in sequential inference. *NIPS*, 2001.

[23] E. Rahm and A. Thor. Citation analysis of database publications. *SIGMOD Record*, 34(4):48–53.

[24] X. Ren, A. El-Kishky, C. Wang, F. Tao, C. R. Voss, and J. Han. Clustype: Effective entity recognition and typing by relation phrase-based clustering. In *KDD*, 2015.

[25] X. Ren, J. Liu, X. Yu, U. Khandelwal, Q. Gu, L. Wang, and J. Han. Cluscite: Effective citation recommendation by information network-based clustering. In *KDD*, 2014.

[26] D. Shen, J.-T. Sun, H. Li, Q. Yang, and Z. Chen. Document summarization using conditional random fields. In *IJCAI*, volume 7, pages 2862–2867, 2007.

[27] W. Shen, J. Wang, P. Luo, and M. Wang. A graph-based approach for ontology population with named entities. In *CIKM*, 2012.

[28] S. Shi, H. Zhang, X. Yuan, and J.-R. Wen. Corpus-based semantic class mining: distributional vs. pattern-based approaches. In *COLING*, 2010.

[29] P. P. Talukdar and F. Pereira. Experiments in graph-based semi-supervised learning methods for class-instance acquisition. In *ACL*, 2010.

[30] Y. Tateisi, Y. Shidahara, Y. Miyao, and A. Aizawa. Annotation of computer science papers for semantic relation extraction. In *IREC*, 2014.

[31] C.-T. Tsai, G. Kundu, and D. Roth. Concept-based analysis of scientific literature. In *CIKM*, 2013.

[32] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning. Kea: Practical automatic keyphrase extraction. In *JCDL*, 1999.

[33] D. Zhou, O. Bousquet, and J. Weston. Learning with local and global consistency. *NIPS*, 2004.

[34] X. Zhu, J. Lafferty, and R. Rosenfeld. *Semi-supervised learning with graphs*. Carnegie Mellon University, Language Technologies Institute, School of Computer Science, 2005.