Организация и управление личными каталогами научных публикаций с использованием технологий Semantic Web

А.А. Бездушный, andrey.bezdushny@gmail.com

Московский физико-технический институт, г. Долгопрудный

Аннотация

В своей повседневной деятельности, человек сталкивается со все большими объемами информации, значительная часть которой хранится в цифровом формате. Задачей организации и ведения данных сведений занимаются системы управления личной информацией. В предыдущих работах была спроектирована базовая архитектура системы семантического управления личной информацией. В данной работе рассматривается модуль данной системы, поддерживающий работу с научными публикациями. Модуль предоставляет пользовательские интерфейсы для организации публикаций, позволяет автоматически выделять метаданные из текстов публикаций, а также загружает дополнительные сведения о публикациях из LOD.

1 Введение

Вопросы, ведения и организации научных публикаций, рассматриваемые в данной работе, являются частью более широкой задачи *управления личной информацией*. Системы управления личной информацией автоматизируют процессы ведения и работы с информационным пространством – совокупностью всех сведений, с которыми человек работает в настоящее время или работал ранее. В предыдущих работах [1], [2] была спроектирована система управления личной информации, модуль которой описывается в данной статье.

В работе любого ученого одной из наиболее важных и затратных по времени задач является поиск и изучение опубликованных ранее работ, поэтому вопросы организации и ведения научных публикаций, чрезвычайно важны. В последние годы, в данном направлении был проведен ряд исследований, которые можно разделить на несколько категорий: исследования в области ведения и обмена научными публикациями, организация систем, объединяющих публикации из различных источников, организация системы, представляющих доступ к публикациям в соответствии со стандартом Linked Open Data (LOD) [3]. В данной работе, в рамках подзадачи управления личной информацией, рассматриваются вопросы управления научными публикациями. К проектируемому модулю предъявляются следующие требования:

- загрузка в систему научных публикаций, хранящихся на компьютере пользователя
- автоматическое выделение метаданных (название, авторы, аннотация) из текстов загруженных публикаций

- организация и ведение сводного репозитория публикаций, полученных из внешних источников
- предоставление пользователю метаданных о загруженных им публикациях, в случае если они были найдены в сводном репозитории

В дальнейшем каждое из требований будет рассмотрено более детально.

2 Существующие решения

Рассмотрим основные решения, которые, так или иначе, касаются задач, описанных в предыдущем разделе.

Первым направлением исследований являются системы, поддерживающие ведение и обмен научными публикациями, такие как BibSonomy [4], CiteULike[1], Mendeley [2]. К основным задачам этих систем относятся ведение каталога публикаций, загрузка и выгрузка публикаций в формате BibTeX, автоматическое выделение метаданных из текстов загруженных публикаций. Описанные задачи во многом пересекаются с требованиями, поставленными в данной работе, но есть и ряд отличий от разрабатываемого решения. Первым отличием является итоговая цель работы – данный модуль разрабатывается как часть системы управления личной информацией, обеспечивающей работу с различными видами данных (email сообщениями, контактами, сведениями из календарей), тогда как представленные решения оперируют только с библиографической информацией. Другим отличием является поддержка возможности получения метаданных о публикациях из внешних источников.

Ко второму направлению исследований можно отнести системы, объединяющие сведения о публикациях, размещенных в различных источниках – CiteSeerX [5], DBLP [6], arXiv[3], IEEE Xplore[4], Google Scholar[5], Microsoft Academic Search[6]. Среди этих решений можно выделить 3 основных категории: системы, в которых сведения о публикациях заносятся вручную (arXiv, IEEE), системы, агрегирующие сведения из заранее выбранных источников (CiteSeerX, DBLP), а также системы, агрегирующие публикации из всей сети (Google Scholar, Microsoft Academic Search).

Третьим направлением исследований является создание протоколов доступа к библиографическим сведениям. Наиболее значимым результатом здесь являются стандарты, формализующие работу открытых архивов публикаций – OAI-PMH [7] и OAI-ORE [8].

В последнем направлении исследуются вопросы использования стандартов Semantic Web и Linked Open Data (LOD) при работе с научными публикациями. В работах [9] и [10] рассматривается задача представления библиографических сведения в RDF формате, а также исследуются возможности для публикации этих сведений в LOD. Также вопросы публикации библиографических сведений в LOD рассматриваются в работе RKBExplorer

[1] http://www.citeulike.org/
[2] http://www.mendeley.com/
[3] http://arxiv.org/
[4] http://ieeexplore.ieee.org/
[5] http://scholar.google.ru/
[6] http://academic.research.microsoft.com/

[11] и системе DBLP[7]. В рамках работ RKBExplorer были созданы RDF выгрузки библиографических сведений имеющихся в сети. В качестве источников этих сведений используются как централизованные онлайн-репозитории публикаций (CiteSeerX, ACM), так и различные открытые архивы, предоставляющие данные по протоколу OAI-PMH.

3 Загрузка и извлечение метаданных из публикаций

Для загрузки публикаций, хранящихся на компьютере пользователя, используется прототип системы управления личной информации, реализованный в рамках работ [1], [2]. Прототип состоит из двух модулей – приложение, отвечающее за перенос файлов с компьютера пользователя в систему и веб портал предоставляющий интерфейс работы с загруженными данными.

Следующим этапом, после загрузки публикаций в систему, является автоматическое выделение метаданных из них. Из текстов публикаций можно автоматически выделить такие метаданные как, название, авторы, аннотация, список литературы. Данная задача относится к классу *задач разметки последовательностей* (sequence labeling). Основной целью этих задач является присвоение ярлыков элементам последовательности наблюдений. При решении задачи выделения метаданных из текстов публикаций, такими наблюдениями являются строки текста, а ярлыками – типы строк (название статьи, аннотация, название разделов). На рисунке 1 приведен пример разбиения текста публикации программой ParsCit. Разными цветами отмечены различные категории строк – название статьи, авторы, место работы авторов, название разделов, содержание разделов.

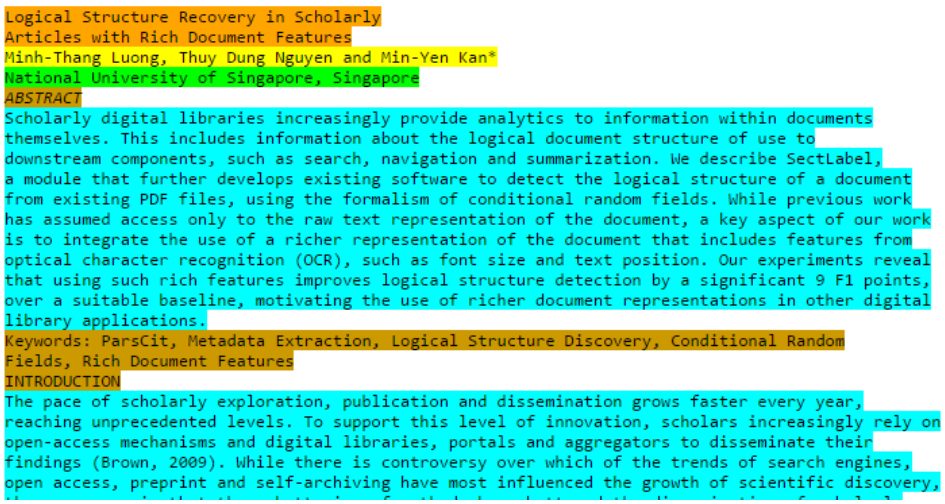
Logical Structure Recovery in Scholarly
Articles with Rich Document Features
Minh-Thang Luong, Thuy Dung Nguyen and Min-Yen Kan*
National University of Singapore, Singapore
ABSTRACT
Scholarly digital libraries increasingly provide analytics to information within documents themselves. This includes information about the logical document structure of use to downstream components, such as search, navigation and summarization. We describe SectLabel, a module that further develops existing software to detect the logical structure of a document from existing PDF files, using the formalism of conditional random fields. While previous work has assumed access only to the raw text representation of the document, a key aspect of our work is to integrate the use of a richer representation of the document that includes features from optical character recognition (OCR), such as font size and text position. Our experiments reveal that using such rich features improves logical structure detection by a significant 9 F1 points, over a suitable baseline, motivating the use of richer document representations in other digital library applications.
Keywords: ParsCit, Metadata Extraction, Logical Structure Discovery, Conditional Random Fields, Rich Document Features
INTRODUCTION
The pace of scholarly exploration, publication and dissemination grows faster every year, reaching unprecedented levels. To support this level of innovation, scholars increasingly rely on open-access mechanisms and digital libraries, portals and aggregators to disseminate their findings (Brown, 2009). While there is controversy over which of the trends of search engines, open access, preprint and self-archiving have most influenced the growth of scientific discovery,

Рис 1. Результат разбиения текста публикации программой ParsCit

Рассмотрим основные подходы к решению поставленной задачи. Можно выделить несколько категорий алгоритмов, используемых для этого: основанные на базе регулярных выражений, основанные на наборе правил поиска, основанные на методах машинного обучения. Наилучшие результаты, на данный момент, показывают решения, основанные на машинном обучении. Среди них наиболее распространены следующие алгоритмы – скрытая марковская модель (Hidden Markov Model, HMM) [12], метод опорных векторов (Support

[7] D2R Server publishing the DBLP Bibliography Database - http://dblp.l3s.de/d2r/

Vector Machine, SVM) [13], модель условных случайных полей (Conditional Random Fields, CRF) [14], [15], [16], [17]. В соответствии с последними исследованиями [18], наибольшей эффективностью среди них обладает модель CRF.

Модель CRF является расширением модели HMM. Отличительной особенностью CRF, является то, что с помощью нее можно получить хорошие результаты даже при наличии тесных взаимосвязей между наблюдаемыми переменными. Рассмотрим принцип работы CRF на примере выделения метаданных из текста научной публикации. Пусть $L = \{l_1, l_2, l_3, \dots\}$ – множество строк в документе, а $C = \{c_1, c_2, c_3, \dots\}$ – множество возможных категорий строк (название, автор, аннотация, итп). Условным случайным полем называется следующая условная вероятность:

$$p(y|x, \lambda) = \frac{1}{Z(x)} * \exp(\sum_j \lambda_j * f_j(y, x))$$

Где $y \in C$, $x \in L$, $f_j(y, x)$ – функции-признаки (feature functions), λ_j – вес j-го признака, рассчитанный в ходе обучения, $Z(x)$ – коэффицент нормализации. В задаче определения типов строк в научных публикациях, часто используются следующие признаки: количество слов в строке, положение строки в документе, вхождение в строку символов, специфичных для названий разделов (например, 1, 1.1, 1.1.1). Вкратце, алгоритм работы CRF можно описать следующим образом:

- во время обучения, для каждого элемента выборки рассчитываются значения функций-признаков, на основании которых определяется вес каждого признака
- при использовании на реальных данных, на основании значения функций-признаков и подсчитанных при обучении весов этих признаков, вычисляется вероятности принадлежности наблюдения к категории

Среди реализаций описанного подхода, обладающих открытым исходным кодом, выделяются работы *ParsCit* [14][15] и *PaperCut* [17]. Требование к открытости исходного кода возникает из необходимости переобучения систем для поддержки российских публикаций. Из этих систем была выбрана система ParsCit, поскольку на данный момент она достаточно продолжительное время успешно используется в онлайн репозитории публикаций CiteSeerX.

Для того чтобы поддержать выделение метаданных из российских публикаций, к обучающей выборке, созданной авторами системы ParsCit, были добавлены сорок публикаций из различных российских конференций. При подборе статей упор делался на максимальное различие в структуре разделов и оформлении. Помимо обновления обучающей выгрузки также потребовалось внести некоторые изменения в исходные коды системы. Тестирование корректности функционирования системы после переобучения проводилось на наборе из тридцати публикаций. В таблице 1 приведены количества публикаций из пробной выборки, для которых удалось выделить те или иные метаданные.

Название	Авторы	Место работы авторов	Аннотация	Список литературы
27/30	20/30	19/30	23/30	28/30

Таблица 1. Результаты работы ParsCit на пробной выборке российских публикаций

4 Организация сводного RDF репозитория

Одним из требований, предъявленных к разрабатываемому модулю, является возможность получения метаданных о публикациях из внешних источников. Самым простым способом достижения данной цели, может показаться последовательный поиск запрошенной публикации в различных источниках. Однако, поскольку количество таких источников велико (только в реестре открытых архивов ROAR[8] на данный момент зарегистрировано 3830 источников), данный вариант является не осуществимым. В связи с этим, возникает потребность организации и ведения сводного репозитория публикаций, объединяющего данные из различных источников. Данный репозиторий в дальнейшем будет использовать для получения метаданных по публикациям, загруженным пользователем. В качестве формата хранения данных в репозитории логично использовать RDF, т.к. это упрощает получение сведений, опубликованных в LOD.

4.1 Загрузка сведений из репозиториев LOD

На данный момент достаточно большой объем научных публикаций, доступных в открытом доступе опубликован в LOD. Рассмотрим основные источники публикаций в LOD, которые имеются на данный момент. В рамках проекта DBLP был запущен D2R[9] сервер, представляющий данные репозитория DBLP в RDF формате и предоставляющий SPARQL точку доступа к ним. Другим источником публикаций являются RDF выгрузки, созданные в рамках работ RKBExplorer. Последние включают в себя выгрузки из различных онлайн репозиториев, таких как CiteSeerX[10], IEEE[11], ACM[12], а также объединенную RDF выгрузку из источников, предоставляющих данные по протоколу OAI-PMH[13].

Описанные RDF выгрузки были положены в основу организуемого сводного репозитория. В качестве базы данных была выбрана RDF СУБД OpenLink Virtuoso. В ходе импорта в сводный репозиторий было загружено 177 миллионов RDF троек и около 9 миллионов публикаций.

В выгрузках, созданных в рамках работ RKBExplorer, используются онтологии http://www.aktors.org/ontology/portal# и http://www.aktors.org/ontology/support#. На рисунке 2 приведен пример ресурса выгруженного из репозитория IEEE Xplore.

[8] http://roar.eprints.org/
[9] http://dblp.l3s.de/d2r/
[10] http://citeseer.rkbexplorer.com/
[11] http://ieee.rkbexplorer.com/
[12] http://acm.rkbexplorer.com/
[13] http://oai.rkbexplorer.com/

```
<rdf:Description rdf:about="http://ieee.rkbexplorer.com/id/publication-c3c2c6df6c287f226afeb673b5d60ba4">
    <akt:has-date>
        <support:Calendar-Date rdf:about='http://www.aktors.org/ontology/date#1971'>
            <support:year-of>1971</support:year-of>
            <support:has-pretty-name>1971</support:has-pretty-name>
        </support:Calendar-Date>
    </akt:has-date>
    <akt:has-title>On the Design of Minimum Length Fault Tests for Combinational Circuits</akt:has-title>
    <rdf:type rdf:resource="&akt;Proceedings-Paper-Reference" />
    <akt:has-author>
        <akt:Person rdf:about="http://ieee.rkbexplorer.com/id/person-15e246ba19026defe842b5662c33f127-c3c2c6df6c287f226afeb673b5d60ba4">
            <akt:full-name>L.W. Bearnson</akt:full-name>
        </akt:Person>
    </akt:has-author>
    <akt:has-author>
        <akt:Person rdf:about="http://ieee.rkbexplorer.com/id/person-8e8d7e2465398a124b379507ad54a93d-c3c2c6df6c287f226afeb673b5d60ba4">
            <akt:full-name>C. C. Carroll</akt:full-name>
        </akt:Person>
    </akt:has-author>
    <akt:paper-in-proceedings>
        <akt:Conference-Proceedings-Reference rdf:about="http://ieee.rkbexplorer.com/id/proceedings-583c8df60227d1be4e191eaf50d8c5e5" >
            <akt:has-title>International Symposium on Fault Tolerant Computing, 1971</akt:has-title>
            <akt:has-date>
                <support:Calendar-Date rdf:about='http://www.aktors.org/ontology/date#1971'>
                    <support:year-of>1971</support:year-of>
                    <support:has-pretty-name>1971</support:has-pretty-name>
                </support:Calendar-Date>
            </akt:has-date>
        </akt:Conference-Proceedings-Reference>
    </akt:paper-in-proceedings>
</rdf:Description>
```

Рис 2. Пример RDF ресурса выгруженного из RKBExplorer

В выгрузках DBLP используются более распространенные онтологии, такие как http://xmlns.com/foaf/0.1/, http://purl.org/dc/terms/. На рисунке 3 приведен пример ресурса из репозитория DBLP.

```
<rdf:Description rdf:about="http://dblp.l3s.de/d2r/data/publications/journals/jsyml/Langford41a">
    <dc:license rdf:resource="http://www.informatik.uni-trier.de/~ley/db/copyright.html" />
    <rdfs:label>RDF Description of List of Officers and Members of the Association for Symbolic Logic.</rdfs:label>
    <foaf:primaryTopic>
        <rdf:Description rdf:about="http://dblp.l3s.de/d2r/resource/publications/journals/jsyml/Langford41a">
            <rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Document"/>

            <dc:title rdf:datatype="http://www.w3.org/2001/XMLSchema#string">List of Officers and Members of the Association for Symbolic Logic.</dc:title>
            <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">List of Officers and Members of the Association for Symbolic Logic.</rdfs:label>

            <dc:creator rdf:resource="http://dblp.l3s.de/d2r/resource/authors/C._H._Langford" />
            <foaf:maker rdf:resource="http://dblp.l3s.de/d2r/resource/authors/C._H._Langford" />
            <foaf:homepage rdf:resource="http://dx.doi.org/10.1017/S0022481200106486" />

            <dc:identifier rdf:datatype="http://www.w3.org/2001/XMLSchema#string">DBLP journals/jsyml/Langford41a</dc:identifier>
            <dc:identifier rdf:datatype="http://www.w3.org/2001/XMLSchema#string">DOI 10.1017%2FS0022481200106486</dc:identifier>

            <dc:bibliographicCitation>
                <rdf:Description rdf:about="http://dblp.uni-trier.de/rec/bibtex/journals/jsyml/Langford41a"/>
            </dc:bibliographicCitation>

            <dc:issued rdf:datatype="http://www.w3.org/2001/XMLSchema#gYear">1941</dc:issued>
            <ns0:volume rdf:datatype="http://www.w3.org/2001/XMLSchema#string">6</ns0:volume>
            <ns0:pages rdf:datatype="http://www.w3.org/2001/XMLSchema#string">174-178</ns0:pages>
            <ns0:number rdf:datatype="http://www.w3.org/2001/XMLSchema#string">4</ns0:number>
            <ns0:journal rdf:resource="http://dblp.l3s.de/d2r/resource/journals/jsyml" />
        </rdf:Description>
    </foaf:primaryTopic>

</rdf:Description>
```

Рис 3. Пример RDF ресурса выгруженного из DBLP

В сводном репозитории было решено следовать подходу, примененному в DBLP – использовать общеизвестные словари там, где это возможно. Поэтому, перед импортом, сведения из RKBExplorer были приведены к форматам, используемым в DBLP.

4.2 Использование источников, не имеющих выгрузок в LOD

Не смотря на то, что довольно большой объем публикаций в структурированном виде доступен в LOD, часть онлайн репозиториев, либо не предоставляют никакой возможности для автоматического получения данных из них, либо предоставляют только API для осуществления поиска. Поэтому, если поиск по сводному репозиторию не дал результатов, системой осуществляется поиск статьи в тех онлайн репозиториях, которые предоставляют

публичное API. В рамках реализованного модуля осуществляется поиск по репозиториям arXiv[14], IEEE Xplore[15], Springer[16]. В случае успешного поиска, найденные сведения приводятся к RDF виду и заносятся в сводный репозиторий.

5 Заключение

В работе был представлен модуль работы с научными публикациями. Основными задачами модуля являются: загрузка в систему научных публикаций с компьютера пользователя, автоматическое выделение метаданных из текстов загруженных публикаций, организация и ведение единого репозитория публикаций. Рассмотрены существующие решения, поддерживающие автоматическое выделение метаданных, среди которых, для дальнейшего использования, выбрано решение ParsCit. Система ParsCit была переобучена на выборке, содержащей российские публикации. Для получения дополнительных метаданных о статьях был организован сводный RDF репозиторий публикаций, содержащий сведения о научных публикациях доступных в LOD.

Перспективным направлением для дальнейших исследований является анализ организованного репозитория. Проводя анализ метаданных публикаций, загруженных в систему, можно осуществлять поиск других, потенциально интересных пользователю публикаций, хранящихся в сводном репозитории.

6 Литература

1 *Бездушный А. А., Бездушный А. Н., Серебряков В. А.* Модель семантического управления личной информацией // Труды 16-й Всероссийской научной конференции «Электронные библиотеки
перспективные методы и технологии, электронные коллекции» - RCDL 2014, Дубна, Россия, 2014.

2 *Бездушный А. А.* Прототипирование системы семантического управления персональной информацией // Труды 56-й Научной Конференции МФТИ «Современные проблемы фундаментальных и прикладных наук», Долгопрудный, Россия, 2013

3 *Bernadette Hyland, Ghislain Atemezing, Boris Villazón-Terrazas.* Best Practices for Publishing Linked Data // W3C recommendation. 2014. [HTML] (http://www.w3.org/TR/ld-bp/)

4 *Hotho A. et al.* BibSonomy: A social bookmark and publication sharing system // Proceedings of the Conceptual Structures Tool Interoperability Workshop at the 14th International Conference on Conceptual Structures. – 2006. – P. 87-102.

[14] http://arxiv.org/help/api/index
[15] http://ieeexplore.ieee.org/gateway/
[16] http://dev.springer.com/

5 *Li H. et al.* CiteSeerx: an architecture and web service design for an academic document search engine // Proceedings of the 15th international conference on World Wide Web. – ACM, 2006. – P. 883-884.

6 *Ley M.* The DBLP computer science bibliography: Evolution, research issues, perspectives //String Processing and Information Retrieval. – Springer Berlin Heidelberg, 2002. – P. 1-10.

7 *Carl Lagoze, Herbert Van de Sompel, Michael Nelson et al.* The Open Archives Initiative Protocol for Metadata Harvesting [HTML] (http://www.openarchives.org/OAI/openarchivesprotocol.html)

8 *Carl Lagoze, Herbert Van de Sompel, Pete Johnston et al.* Open Archives Initiative Object Reuse and Exchange [HTML] (http://www.openarchives.org/ore/1.0/primer)

9 *Xin R. S. et al.* Publishing bibliographic data on the Semantic Web using BibBase // Semantic Web Journal. – 2013. – V. 4. – №. 1. – P. 15-22.

10 *Haslhofer B., Schandl B.* Interweaving OAI-PMH data sources with the linked data cloud // International Journal of Metadata, Semantics and Ontologies. – 2010. – V. 5. – №. 1. – P. 17-31.

11 *Glaser H., Millard I. C., Jaffri A.* RKBExplorer.com: a knowledge driven infrastructure for linked data providers. // Proceedings of European Semantic Web Conference (ESWC), Tenerife, Spain - 2008 - P. 797–801

12 *Takasu A.* Bibliographic attribute extraction from erroneous references based on a statistical model // Proceedings of the Third ACM/IEEE-CS Joint Conference on Digital Libraries, ACM Press, New York – 2003. – P. 49-60.

13 *Han H. et al.* Automatic document metadata extraction using support vector machines // Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital libraries, ACM Press, New York – 2003. – P. 37-48.

14 *Councill I. G., Giles C. L., Kan M. Y.* ParsCit: an Open-source CRF Reference String Parsing Package // Proceedings of the Sixth International Language Resources and Evaluation (LREC'08). Marrakech, Morocco: European Language Resources Association (ELRA). – 2008. – V. 1. – P. 661–667

15 *Luong M. T., Nguyen T. D., Kan M. Y.* Logical structure recovery in scholarly articles with rich document features // International Journal of Digital Library Systems (IJDLS). – 2010. – V. 1. – №. 4. – P. 1-23.

16 *Peng F., McCallum A.* Accurate information extraction from research papers using conditional random fields // Proceedings of Human Language Technology Conference and North American Chapter of the Association for Computational Linguistics (HLT-NAACL'04), Boston, MA – 2004. – V. 1. – P. 329–336.

17 *Candeias R., Calado P., Martins B.* Metadata Extraction from Scholarly Articles Using Stacked CRFs. [PDF] (http://papercut.googlecode.com/hg-history/98464ac0efb47c55159b313c89b0b305ba1d83f9/PaperCutTesting/targetPDF/success/papercut.pdf)

18 *Granitzer M. et al.* A comparison of layout based bibliographic metadata extraction techniques // Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics. – ACM, 2012. – V.1 – P. 19.