BIOINFORMATICS TOOL RECOMMENDATION BASED ON USAGE CONTEXT EVIDENCES

ANGKANA HUANG

**A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE (COMPUTER SCIENCE)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2017**

Thesis
entitled

BIOINFORMATICS TOOL RECOMMENDATION BASED ON USAGE CONTEXT EVIDENCES

.....
Miss. Angkana Huang
Candidate

.....
Lect. Apirak Hoonlor,
Ph.D. (Computer Science)
Major advisor

.....
Prof. Peter Haddawy,
Ph.D. (Computer Science)
Co-advisor

.....
MAJ Damon W. Ellison,
Ph.D. (Molecular Microbiology and
Microbial Pathogenesis)
Co-advisor

.....
Prof. Patcharee Lertrit,
M.D., Ph.D. (Biochemistry)
Dean
Faculty of Graduate Studies
Mahidol University

.....
Asst. Prof. Boonsit Yimwadsana,
Ph.D. (Electrical Engineering)
Program Director
Master of Science Programme
in Computer Science
Faculty of Information and
Communication Technology
Mahidol University

Thesis
entitled

BIOINFORMATICS TOOL RECOMMENDATION BASED ON USAGE CONTEXT EVIDENCES

was submitted to the Faculty of Graduate Studies, Mahidol University
for the degree of Master of Science (Computer Science)
on
January 5, 2017

.....
Miss. Angkana Huang
Candidate

.....
Asst. Prof. Thanawin Rakthanmanon,
Ph.D. (Computer Science)
Chair

.....
Lect. Apirak Hoonlor,
Ph.D. (Computer Science)
Member

.....
MAJ Damon W. Ellison,
Ph.D. (Molecular Microbiology and
Microbial Pathogenesis)
Member

.....
Prof. Peter Haddawy,
Ph.D. (Computer Science)
Member

.....
Prof. Patcharee Lertrit,
M.D., Ph.D. (Biochemistry)
Dean
Faculty of Graduate Studies
Mahidol University

.....
Assoc. Prof. Jarernsri L. Mitrpanont,
Ph.D. (Computer Science)
Dean
Faculty of Information and
Communication Technology
Mahidol University

ACKNOWLEDGEMENTS

First and foremost, I would like to express my gratitude to Dr. In-Kyu Yoon who have encouraged my research growths over the years at AFRIMS and set the fire on my study in this degree. His effort was made solid by Dr. Louis R. Macareo. The generous understandings and extensive supports of Dr. Mac and Dr. Damon W. Ellison have enabled this thesis to be completed within a tight timeline. My time in conducting this research could not have been better without Dr. Apirak Hoonlor, my dear advisor. His consistent support, straightforward guidance, and open-minded working style has improved a lot of my research skills and knowledge. All the coffee treats throughout the time should not be forgotten as well. It will be remembered that Ms. Russama and Ms. Thunyathorn had gone many extra miles to push forward the paperwork rushes. The next significant person I would like to thank is Dr. Wiriya Wheeler. Her open arms that invited me into her bioinformatics team have led me to my research topic, one which I highly enjoyed working on. Another person which I must not miss to acknowledge is Dr. Chonticha Klungthong who supported me in every way a senior colleague could. Last but not least, I will remember how understanding and supportive my family have been over such a busy year.

Angkana Huang

BIOINFORMATICS TOOL RECOMMENDATION BASED ON USAGE CONTEXT EVIDENCES

ANGKANA HUANG 5837702 ITCS/M

M.Sc. (COMPUTER SCIENCE)

THESIS ADVISORY COMMITTEE: APIRAK HOONLOR, Ph.D., PETER HADDAWY, Ph.D., DAMON W. ELLISON, Ph.D.

ABSTRACT

The abundance of bioinformatics tools has grown exponentially over the last three decades. Concurrently, many tools become outdated due to their discontinuation. Staying updated is highly difficult and is costly in terms of time and effort. The existing systems to ease tool discovery primarily attempts to index all the available tools according to their functionalities. Some of them provides the number of cites the tools received to indicate their popularity. The size of these lists have grown large calling for more targeted retrieval approaches. Since the tools are typically used in conjunction, a recent approach allows the users to browse the tools according to expert maintained pipelines. However, generating and updating these pipelines remains manual. Moreoever, the actual conjunct use of the tools are not provided. This thesis suggests an automated pipeline derivation method through literature mining and cross-citation analysis. The analysis patterns and the tool usage patterns were recovered from the data. Recommendation models were built from the data and the derived patterns. Through evaluating the models, actual tool selection behaviors were also understood. During 2009-2016, the tool functionalities with their popularity considered was highly predictive of whether they have been chosen. Along the period, substituting the overall popularity with the local popularity within the recovered analysis patterns became increasingly predictive and was on par in 2016. Such results implies that the recovered patterns resemble the pipelines used in community. Lastly, the pipelines were queried into the recommendation models to obtain the community accepted best-practices. These analysis patterns and best-practices can be used to inform experts regarding the status-quo of the field and can be used as guidelines for newcomers entering the field.

KEY WORDS : RECOMMENDATION SYSTEM / INFORMATION RETRIEVAL / BIOINFORMATICS

67 pages

การแนะนำเครื่องมือทางชีวสารสนเทศโดยอิงหลักฐานทางการใช้งาน

BIOINFORMATICS TOOL RECOMMENDATION BASED ON USAGE CONTEXT EVIDENCES

อังคณา หวัง 5837702 ITCS/M

วท.ม. (วิทยาการคอมพิวเตอร์)

คณะกรรมการที่ปรึกษาวิทยานิพนธ์: อภิรักษ์ หุ่นหล่อ, Ph.D., Peter Haddawy, Ph.D., Damon W. Ellison, Ph.D.

บทคัดย่อ

จำนวนเครื่องมือทางชีวสารสนเทศได้ทวีขึ้นมากในสามทศวรรษที่ผ่านมา ในขณะเดียวกันเครื่องมือจำนวนมากก็ได้หยุดการพัฒนาลง การคงซึ่งความทันสมัยจึงเป็นไปโดยยาก ระบบในปัจจุบันมักมุ่งเน้นการทำดัชนีเครื่องมือและประโยชน์ใช้สอยของเครื่องมือนั้น จัดทำเป็นรายการเพื่อให้เอื้อต่อการค้นพบเครื่อง จำนวนเครื่องมือที่มีมากทำให้การแสดงรายการต้องมีความจำเพาะเจาะจงยิ่งขึ้น ด้วยเครื่องมือชีวสารสนเทศมักถูกใช้ร่วมกันอย่างเป็นลำดับขั้น เมื่อไม่นานมานี้ จึงมีระบบหนึ่งแสดงรายการโดยจัดหมวดหมู่เป็นลำดับตามลำดับขั้นการใช้งาน อย่างไรก็ดี ลำดับขั้นเหล่านี้ยังคงต้องถูกกำหนดและปรับเปลี่ยนโดยมนุษย์ การวิจัยนี้จึงมุ่งให้สามารถค้นพบลำดับขั้นการใช้งานที่มีอยู่ในวรรณกรรมโดยอัตโนมัติและลำดับขั้นที่ถูกค้นพบนั้นไปสร้างระบบแนะนำเครื่องมือชีวสารสนเทศ การประเมินประสิทธิภาพการแนะนำเครื่องมือได้ทำให้ทราบว่าในระหว่าง ค.ศ.2009-2016 ประโยชน์ใช้สอยของแต่ละเครื่องมือร่วมกับความนิยมโดยรวมของเครื่องมือนั้น ๆ ส่งผลสูงสุดต่อการเลือกเครื่องมือ ส่วนความนิยมของเครื่องมือที่จำเพาะต่อลำดับขั้นการใช้งานนั้นเริ่มมีอิทธิพลมากขึ้นในระยะหลังและสูงเทียบเท่ากับความนิยมโดยรวมในปีสุดท้ายของการศึกษา ผลการวิจัยนี้แสดงให้เห็นว่าลำดับขั้นการใช้งานที่ค้นพบโดยอัตโนมัตินี้มีความใกล้เคียงกับลำดับขั้นที่แท้จริงในข้อมูล ประโยชน์ใช้สอยที่อยู่ในลำดับขั้นเหล่านั้นถูกป้อนเข้าสู่ระบบแนะนำเครื่องมือที่สร้างขึ้นเพื่อแสดงชุดเครื่องมือที่เป็นที่ยอมรับและทันสมัยในปัจจุบัน ความรู้นี้เป็นประโยชน์ต่อการติดตามความเปลี่ยนแปลงในการใช้เครื่องมือชีวสารสนเทศ

67 หน้า

CONTENTS

CONTENTS (cont.)

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES (cont.)

LIST OF FIGURES (cont.)

CHAPTER I
INTRODUCTION

1.1 Motivations for a tool recommender

Bioinformatics today refers to the use and the engineering of informatics tools to manage and analyze biology related data [3]. Functionalities of these bioinformatics tools evolve to reflect the changes of the research questions in biology. Many times, pre-existing tools are replaced by newer tools with improved algorithms or implementations. Hence, the best practices change temporally. For the last three decades, the number of resources in bioinformatics have grown exponentially. Keeping track of all the available resources, the *resourceome*, has become highly challenging. Selecting among them is also difficult. Inappropriate tool selction can cost from extended computation time due to inefficiency, suboptimal results due to compromised algorithm performances, to incorrect conclusions caused by assumption violations. To avoid such mistakes, large amount literature are reviewed for support. The believe in these evidences varies. More credible sources are more likely to be trusted. Contextually similar contexts are viewed to be more applicable. With sufficient believes, tools are selected as candidates. Unfortunely, the criteria for evidence sufficiency are likely to be relaxed when time is limited. With a tool recommendation model which well mimics the tool selection behaviors in the field, much of the time and effort used in literature reviewing can be saved.

1.2 Research contributions

This research have shown that data of bioinformatics tool usages can be gathered via a cross-citation analysis approach. An iterative graph clustering method developed during this research was able to recover both the analysis patterns and the

tool usage patterns from the data. Multiple tool selection models were studied. Overall, the best performing model was able to recover over 75% of the tools used when limiting the tool recommendation list to ten entries. The tool selection behaviors in the field were inferred by comparing the performances of the models. Such comparisons if done prospectively, can be used to monitor the behavioral changes of bioinformatics tool selection.

To demonstrate how knowledge from this research can be used, the recovered analysis patterns were queried to the best performing model to provide tool recommendations suitable for each analysis pattern. Such recommendations can be useful in guiding bioinformatics researchers with limited experience in the field or can be informative for experienced researchers to keep up-to-date with the changing trends.

1.3 Research outline

So far, a brief explanation of bioinformatics have been given. Chapter 2 provides a more extensive summary of the timeline in bioinformatics since its emergence till today. The history is introduced to provide a comprehensive understanding of the field evolvement and the issues that arised from its growth. The review then covers the currently avaialble options in coping with the resourceome and stresses on what remained unsolved. The computer science methods relevant to this research were explained in Section 2.3. In Chapter 3, the methods of this research is detailed step-by-step from the data gathering, the analytics, to the outcome evaluation. The results are given and discussed in Chapter 4. The last chapter wraps up the whole research and suggests the works to be done in the future.

CHAPTER II
LITERATIVE REVIEW

2.1 Bioinformatics

The word *bioinformatics* can be viewed as a conjunction between bio- referring to biology, the studying of life, and -informatics which concerns with the processing of data. In that view, the broad arena of biology may be assumed all inclusive. More precisely, the biology is often restricted to the quantitative study of biological macromolecules, i.e. deoxyribonucleic acid (DNA), ribonucleic acid (RNA) and proteins [4]. Since a lot of bioinformatics today utilizes computers to achieve their task, it became commonly confused with computational biology. According to [3], it is technically feasible to separate bioinformatics from computational biology; the first refers to the *engineering* of informatics technology in application to biological data; the second designates the *science* of using computation (with or without computers) to study biology. Though these definitions have been set since year 2000, the term bioinformatics has rooted from as far back as 1970 [5]. Back then, the word *bioinformatics* was used to describe the generation-wise biological information transfer from antecedents to descendents and from genetics embedded in the DNA to high level phenotypic expressions in the organism. Today, those processes are complexly being studied using tools developed through bioinformatics under the hood of *systems biology*.

2.1.1 History and Emergence of Bioinformatics

It was less than a century ago when learning the sequence of amino acids in protein was enabled thru the laboratory method published by Sanger and Tuppy in 1951. In 1962, the speed of data generation was accelerated by automating a more refined version of the sequencing procedures and introducing computers to reconstruct these peptide fragments back into their long chains of polypeptide prior digestion [6]. The first database to store and make available those sequences was the Atlas of Protein

Sequence and Structure established in the 1960s which later became online as the Protein Information Resource (PIR) [7]. Having learnt in the 1950s that three-dimensional structure of proteins are governed by the sequence of their amino acids, efforts were put into aligning the sequences to enable their comparison. The computational difficulty in finding the optimal alignment led to the development of one of the first bioinformatics algorithms. The method explores all possible alignments of two sequences and calculate scores to determine the most probable of them [8].

The knowledge that DNA are genetic information carriers was hinted since 1944. By 1950s, its double helix shape and it compositions was already known. Backboned on repeating sugar groups and phosphate groups, the two strands were inner binded by adenine(A)-thymine(T) pairs and guanine(G)-cytosine(C) pairs (Figure 2.1).

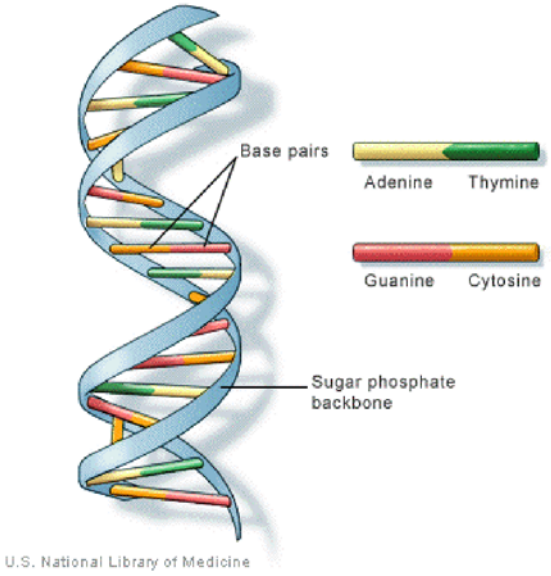


Figure 2.1: An illustration of the DNA double helix structure [1]

Nonetheless, the translation from nucleotide sequences to the synthesis of polypeptides was not certain till 1966 [9] when the triplet code was deciphered and proven with lab experiments. In the findings, linear strings of ATGC bases representing the DNA sequence can be translated into amino acids by decoding three bases of nucleotides (triplet/codon) into one amino acid. Translating each adjacent non-overlapping triplet reveals the sequence of a polypeptide. Around the central dogma ”DNA makes RNA, RNA makes proteins, proteins make us” [10], a new field named *molecular bi-*

ology emerged to study the sophisticated factors and conditions in these translations. With the computational complexity of the problems, computer science fused in. Many more algorithms were developed to solve evolving problems as the body of knowledge in biology grew and shifted [11].

2.1.2 Accelerators of Bioinformatics Growth

Post war in the 1960s, computation powers were transferred to the academics [6]. The complex computations unachievable by hand has moved the frontiers of many fields including biology. The rate of data generation also dramatically increased from the days the first DNA genome was sequenced by Sanger and his group in 1977 with the price inversing falling [12]. Figure 2.2 shows the increasing number of base pairs able to be sequenced per dollar over the years.

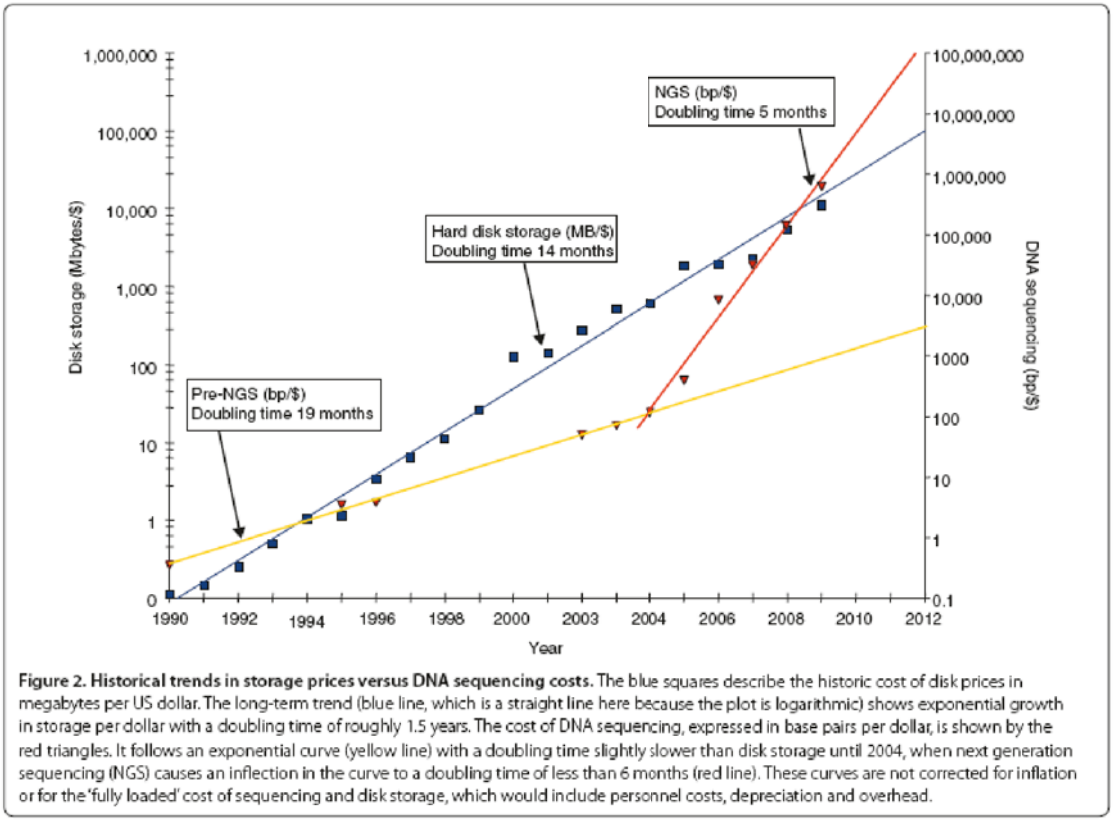


Figure 2. Historical trends in storage prices versus DNA sequencing costs. The blue squares describe the historic cost of disk prices in megabytes per US dollar. The long-term trend (blue line, which is a straight line here because the plot is logarithmic) shows exponential growth in storage per dollar with a doubling time of roughly 1.5 years. The cost of DNA sequencing, expressed in base pairs per dollar, is shown by the red triangles. It follows an exponential curve (yellow line) with a doubling time slightly slower than disk storage until 2004, when next generation sequencing (NGS) causes an inflection in the curve to a doubling time of less than 6 months (red line). These curves are not corrected for inflation or for the 'fully loaded' cost of sequencing and disk storage, which would include personnel costs, depreciation and overhead.

Figure 2.2: Trends of the sequencing cost [2]

Another technology which greatly drived the growth was the Internet. It is needless to convince on how the Internet has changed the human society. Bioinformatics too has been accelerated with the interconnectivity. Data and efforts across the world were linked up and made projects which require world-wide collaborations capable.

2.2 Bioinformatics Today; the Resourceome

Today, bioinformatics is being powered by several big database infrastructures (e.g. GenBank, EMBL) equipped with various ontologies. Data are continuingly being produced and new sequencing technologies are still arising. Retrievals can be made world-wide and analyses can be done either locally or on cloud services. The borders among different fields in biology have become fairly subtle with cross field analyses enabled. As knowledge grew, both breadth and depth of the research questions expanded. New tools were rapidly developed to answer them soundly and effectively. The ability to utilize the diversely developed tools has become a type of expertise, a scarce one. To better utilize the limited human resource, centralized analysis facilities are created and shared among biologist of various fields [13]. While the vast amount of available tools demand indepth understanding of their behavior and context suitability, time and manpower constraints limits these bioinformaticians to perform such compliance. For non-centralized facilities, ones performing bioinformatics analyses are commonly traditional lab biologist which are less likely to have indepth understanding in computer algorithms. Both types of facility structures suffer from keeping up with the whole set of resources available, the *resourceome*. To overcome the phenomena, better indexing and retrieval systems are needed. Challenges to the success were stated in [14], summarized as follow:

– Lack of peer-review in the system
– Lack of useful index for retrieval
– Poor adoption and maintenance of ontologies
– Difficulty in indexing the all tools that exists

2.2.1 Attempts to Address the Resourceome

To address the tool outgrowth, many tool lists were created to index the available resources for retrieval. The lists were made available in the Internet for accessibility. Since 2003, Bioinformatics.ca has been providing manually curated resource links directory with regular annual updates [15]. The website allows browsing the tools through directories. A search box and search filters are provided (Figure 2.3). Each tool is presented with the name, the corresponding website, the associated directory tags, a short description, and a link to the tool's publishing article(s) in PubMed. The number of citations to these publishing articles are given if available (Figure 2.4).

Online Bioinformatics Resources Collection (OBRC) at the University of Pittsburgh Health Sciences Library System provided an alternative search result presentation in 2007. Search results were clustered by Vivisimo, a document clustering engine, to display the search results in a more human-digestable way [16]. Their clustering was based on the contents extracted from the PubMed abstracts and information posted on the corresponding websites. The OBRC website mentioned in the publication is no longer accessible. However, the Vivisimo Clustering Engine is still used today in their Health Sciences Library System. Figure 2.5 shows an example screenshot of the system's result. The 'Topic' tab on the left is derived from the search results clustering.

A web application named BIRI (Figure 2.6) takes a resource's publication abstract as input for automatic resource metadata extraction and indexing [17]. Extending from the capabilities of BIRI, the same group later launched e-MIR2 (Figure 2.7) in 2012 which periodically parses titles and abstracts of new PubMed indexed publications in the medical informatics category to discover new resources. Articles with words matching to predefined classification schemas learnt from expert selected training sets were used to update the inventory list [18]. Alternative to automation, community effort was used to curate the list of resources under the coordination of EXILIR Denmark [19] (Figure 2.8).

Thus far, all of the tool lists have treated each tool entry as an individual entity. OmicTools (Figure 2.9) has introduced a different list presentation approach by taking into account the application and technology specific pipelines [20]. By browsing down the technology or applications, applicable analysis steps are presented. In each

step, candidate tools are listed with performance assessment publications if available.

2.2.2 An Unaddressed Facet of the Resourceome

Either learning the existence of a resource directly from publications or from tool lists, bioinformaticians or biologists today are able to obtain large sets of tool candidates for their work. Therefore, the challenge resides in choosing among the vast number of tools. A logical practice to downsize the number of candidates is through evidence gathering. Tools successfully used by renowned scientists gains better reputations. [21] has shown evidence that *best practices* are field specific and can be indicated from practices of whom they called as *domain authors*, ones who collaborated the most with others. Once sufficient amount of support is found, or time limit is reached, a tool is selected for use. Such tool selection process is highly time consuming. The number of evidences gathered is limited to the time constraints.

2.3 Methods Review

2.3.1 Gathering reputable usage evidences

To build tool recommendation models, data of existing bioinformatics tools and their usage evidences are needed. It is trivial that sources of greater trustworthiness is more probable of being trusted. The scholarly community is an established space to publicize research findings with considerably controlled quality. Though non-homogenous, the reputation of scholarly articles are generally higher than other digitized sources such as websites or blogs. Impacts of these publications have widely been measured through citation analysis. Commonly, these influences are quantified in researcher/author-level, article-level, and journal-level. Using citation counts tracked by citation databases such as Scorpus, PubMed and Google Scholar, many indexes can be calculated. It must be minded that these numbers represent the impact of the cited on the scholar community without considering the citation semantics. The semantics are known to vary among the citers and the subject of interest. Citation context analysis (CCA) framework have been developed to account for these semantics variations [22].

Despite the more accurate relationship representation, the CCA is limited by the accessibility of the citers' full-text articles. Using such approach, Duck et. al. [23] acquired a corpus of 22,376 open-access articles from PubMed Central from 13-years of data. Constraint by the number of articles gathered, analyses were done by combining data of several years. The yearly usage patterns were not recoverable.

2.3.2 Infomap: graph clustering algorithm

Graph clustering is a quite popular field. Common methodologies for graph clustering have been reviewed and published in [24]. Infomap [25] is one of the many clustering algorithms which have been widely used and have shown robust results. The algorithm is based on random walk. Despite its stochastic approach, it has shown to be accurate and fast in clustering graphs with over 10,000 nodes and yield consistent results over multiple runs. Many applications of these graph clustering algorithms have also been published. Recommendation systems are one of those many applications [26, 27].

2.3.3 Unifying the different forms of a word

There are two big genres of word normalization, stemming and lemmatization. While both aims to reduce the different forms of the same base words, stemming is far more crude than lemmatization [28]. In stemming, language-specific rules are used to trim off the ends of the words. The results may not be the correct lemmas of the words. In contrast, lemmatization includes morphological analysis of the words to accurately return the correct lemmas. Both methods of word normalization increases the recall in information retrieval, but at the same time, reduces the precision. The most common stemmer used for English is Porter Stemmer [29]. Other stemming algorithms do exist, e.g. Lovins stemmer and Paice stemmer.

2.3.4 TF-IDF: Term-frequency and inverse document frequency

The term frequency is the number of times a term appears within a document, often divided by the total number of words in the document for normalization. The inverse document frequency is a weight used to offset the term frequency to emphasize words that are more discriminative among the documents. Such weights have

shown to overcome the effects of stop-words [28], word which are typical to most documents and therefore are not beneficial for information retrieval. In summary, the TF-IDF weight is calculated as per Equations 2.1-2.3.

$$\text{TF}(t) = \frac{\text{Number of times t appeared in a document}}{\text{Total number of terms in a document}} \tag{2.1}$$

$$\text{IDF}(t) = \frac{ln(\text{Total number of documents})}{\text{Number of documents with term t in it}} \tag{2.2}$$

$$\text{TF-IDF weight of term t} = \text{TF}(t) * \text{IDF}(t) \tag{2.3}$$

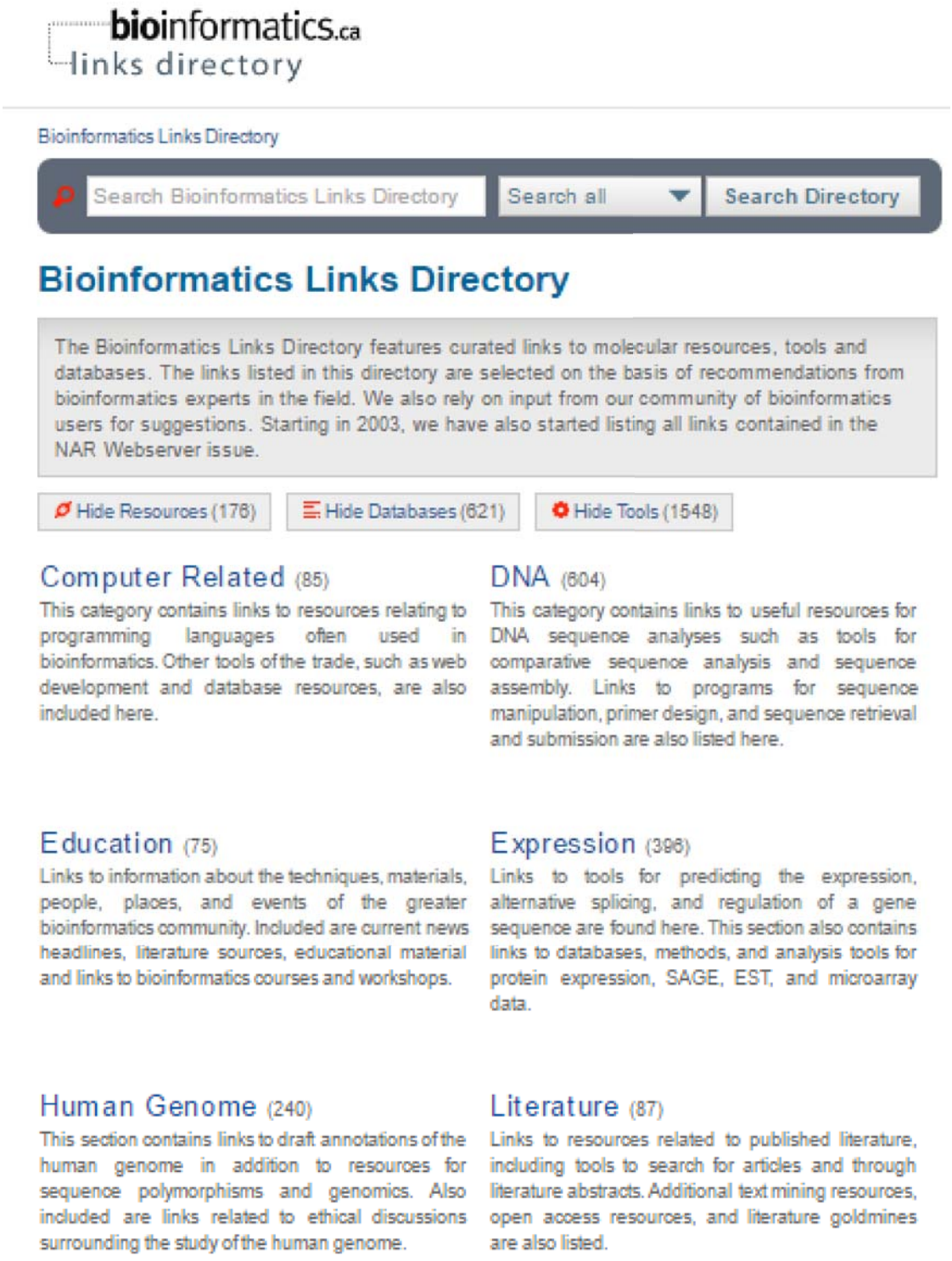


Figure 2.3: A screen shot of Bioinformatics.ca Links Directory

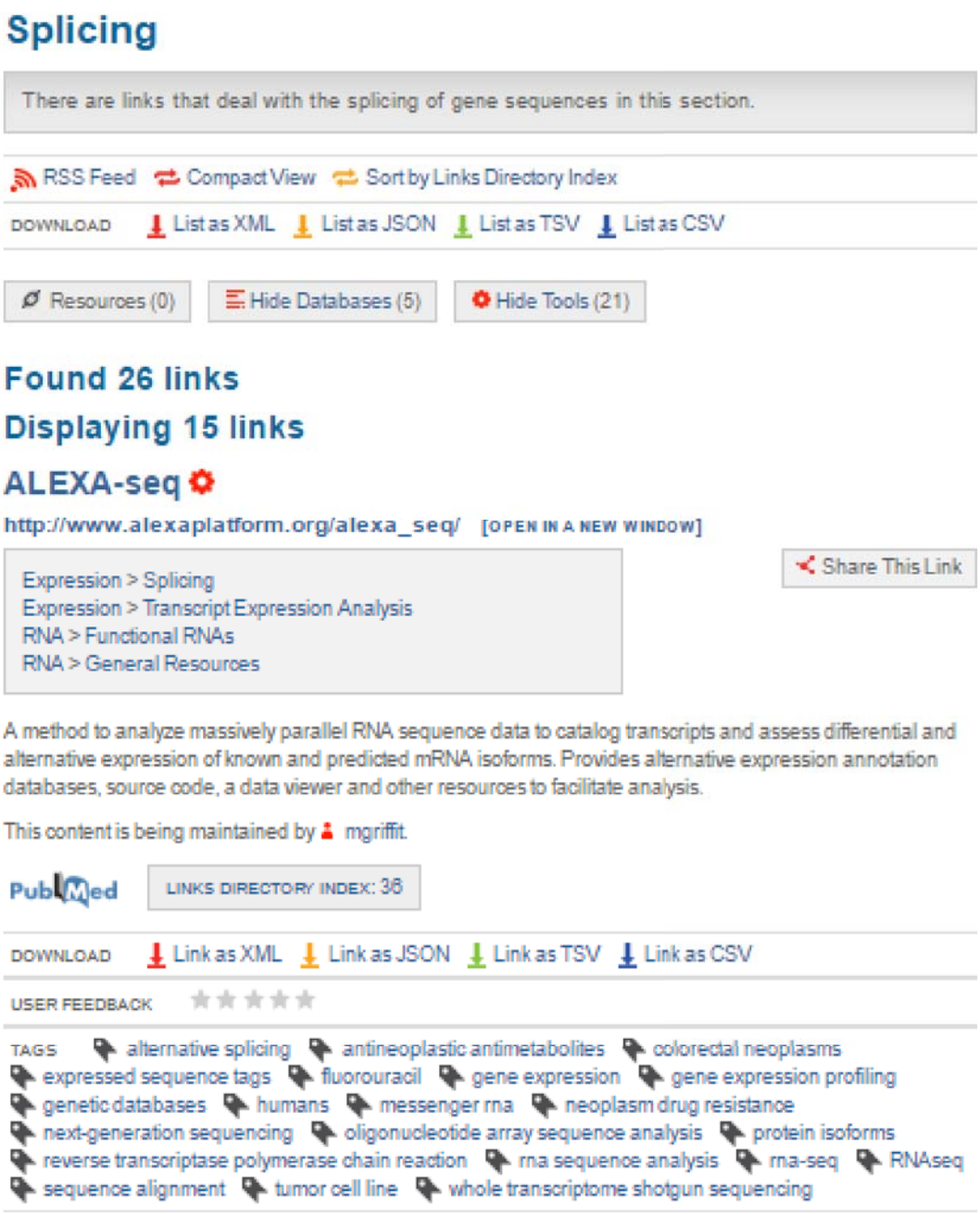


Figure 2.4: An example search result from Bioinformatics.ca Links Directory

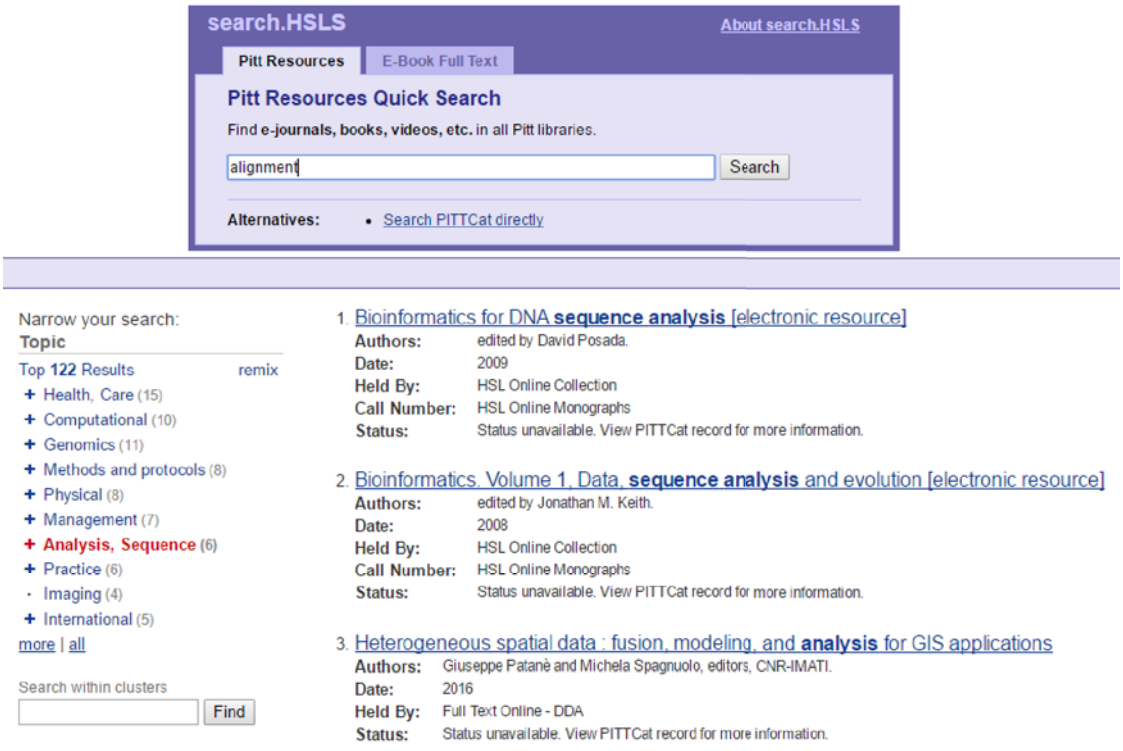
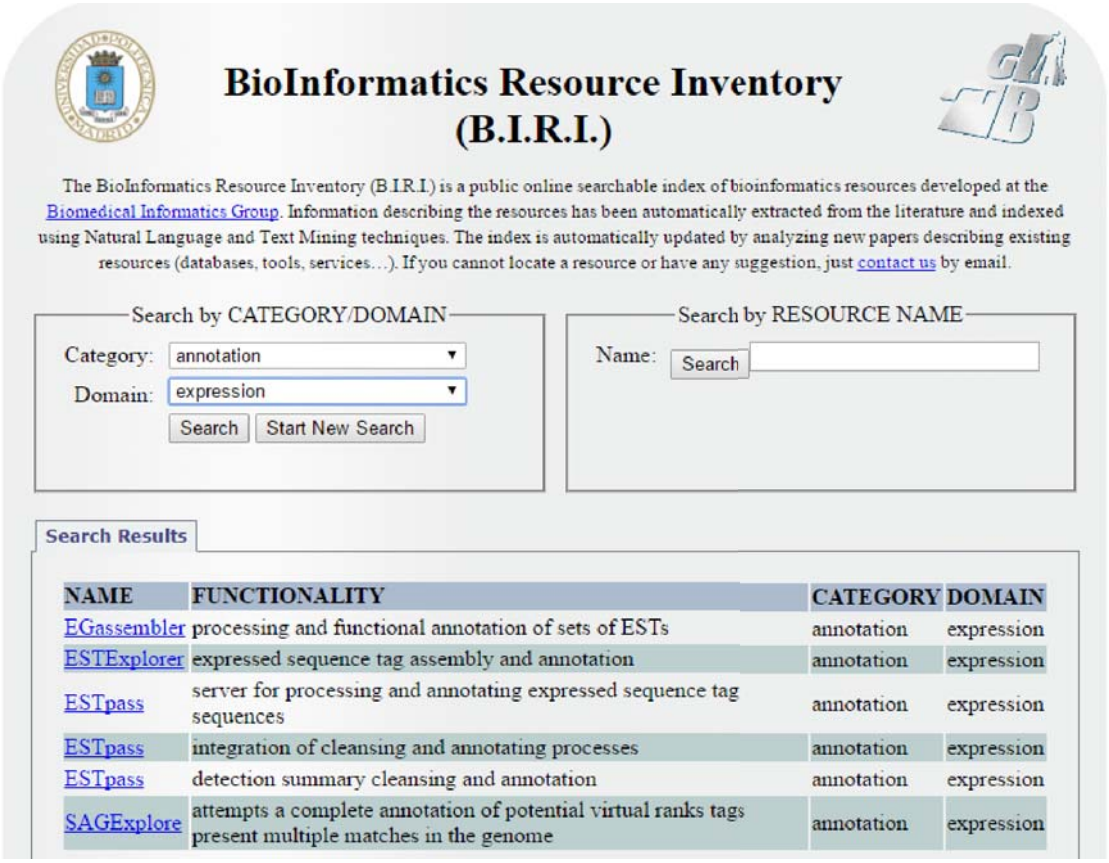


Figure 2.5: An example search result from Health Sciences Library System, Pittsburgh



BioInformatics Resource Inventory (B.I.R.I.)

The BioInformatics Resource Inventory (B.I.R.I.) is a public online searchable index of bioinformatics resources developed at the Biomedical Informatics Group. Information describing the resources has been automatically extracted from the literature and indexed using Natural Language and Text Mining techniques. The index is automatically updated by analyzing new papers describing existing resources (databases, tools, services…). If you cannot locate a resource or have any suggestion, just contact us by email.

Search by CATEGORY/DOMAIN

Category: annotation

Domain: expression

Search | Start New Search

Search by RESOURCE NAME

Name: Search

Search Results

NAME	FUNCTIONALITY	CATEGORY	DOMAIN
EGassembler	processing and functional annotation of sets of ESTs	annotation	expression
ESTExplorer	expressed sequence tag assembly and annotation	annotation	expression
ESTpass	server for processing and annotating expressed sequence tag sequences	annotation	expression
ESTpass	integration of cleansing and annotating processes	annotation	expression
ESTpass	detection summary cleansing and annotation	annotation	expression
SAGExplore	attempts a complete annotation of potential virtual ranks tags present multiple matches in the genome	annotation	expression

Figure 2.6: An example search result from Bioinformatics Resource Inventory (B.I.R.I.)

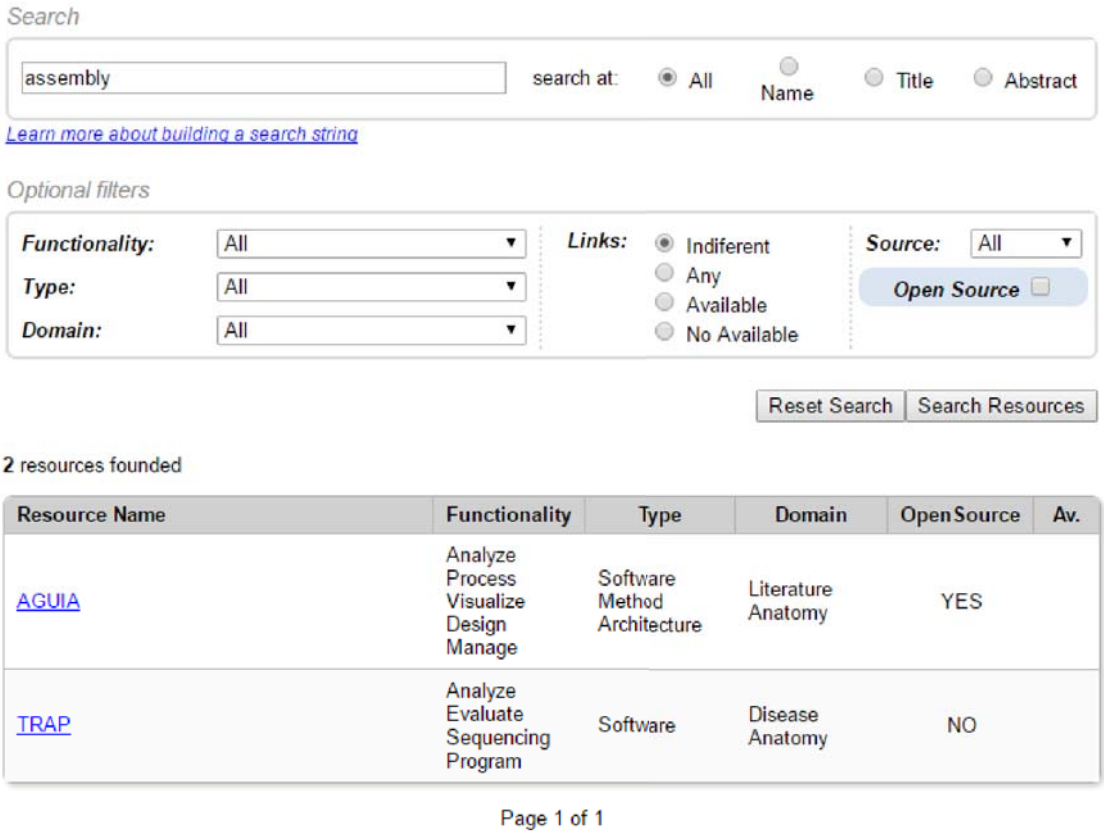


Figure 2.7: An example search result from Electronic Medical Informatics Repository of Resources (e-MIR2)

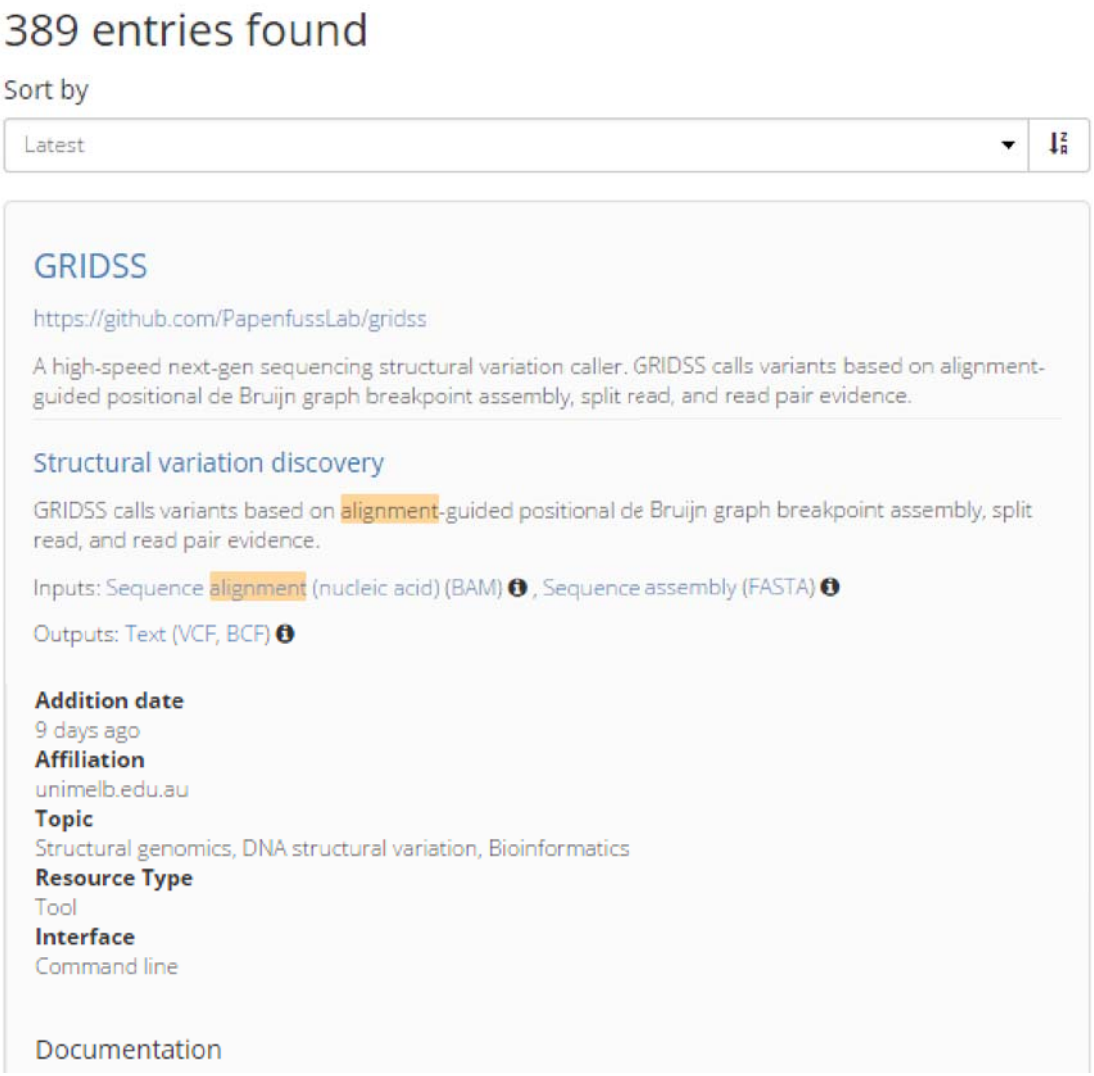


Figure 2.8: An example search result from ELIXIR Tools and Data Services Registry

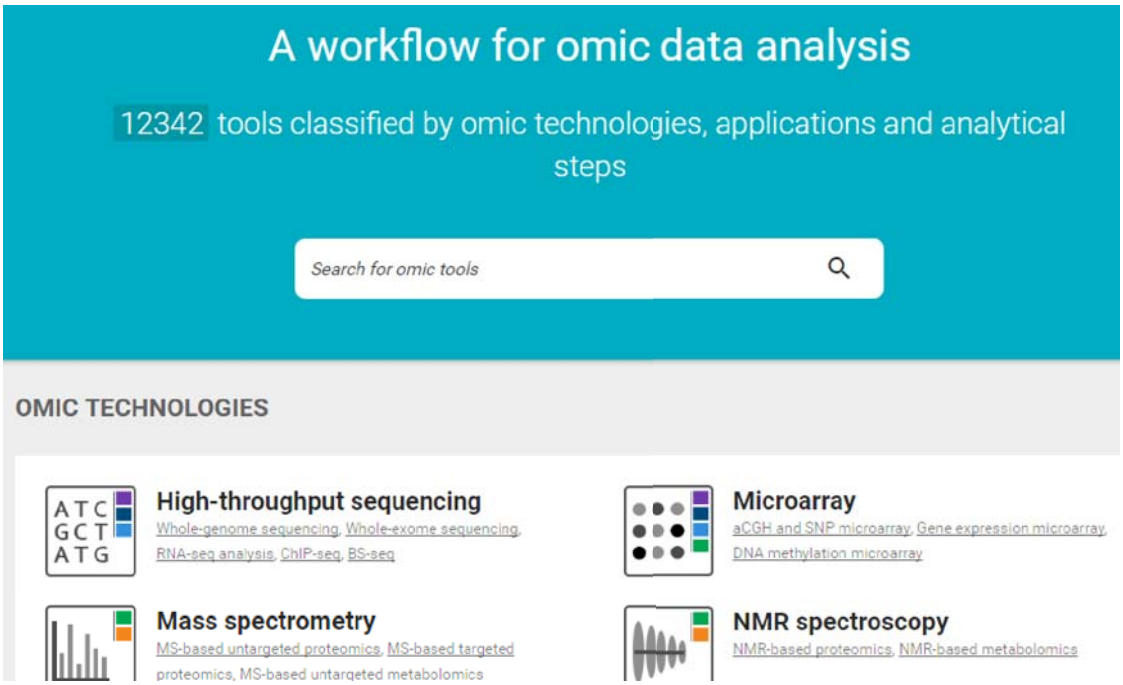


Figure 2.9: An example search result from OmicTools.com

CHAPTER III
METHODOLOGY

3.1 Data retrieval and preprocessing

A finite universe of bioinformatics tool was identified from web parsing two big bioinformatics tool listers, i.e. Bioinformatics.ca and Omicstools.com. Data retrieved from both lists were the tool names, their corresponding websites, PubMed unique identifiers (PMID) of their publications and functionalities of the tools. 1536 tools were identified from Bioinformatics.ca and 4687 tools from Omictools.com. The tools identified from both lists were matched when meeting one of the following criteria:

1. If both website and PMID is not available in one of the lists, match only on identical tool names

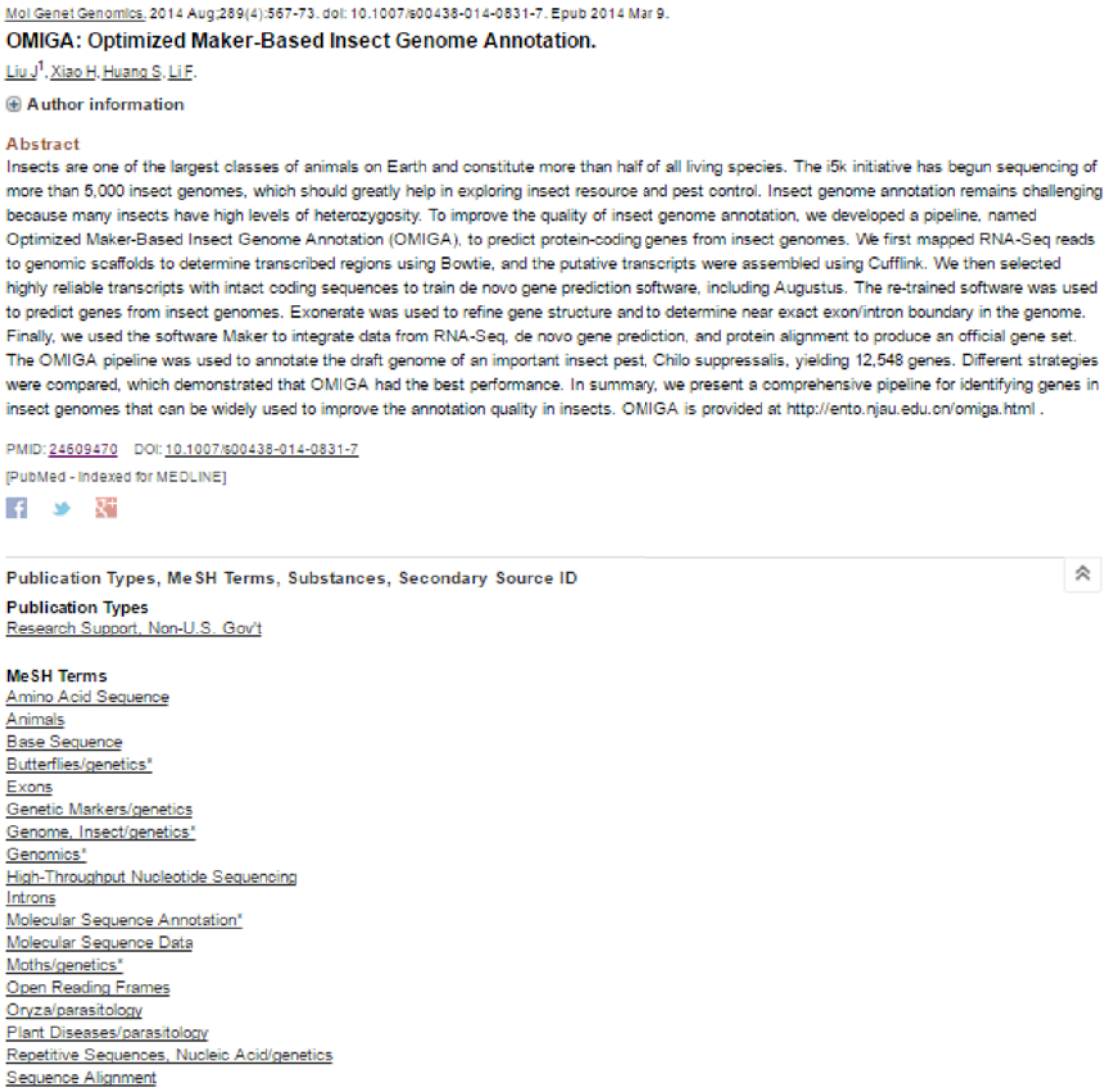
2. If website and/or PMID is available for both lists, match if two of the three elements (name, PMID, URL) match.

3. *To account for minor URL variabilities, the URLs were considered matched if over 80% of the tokenized URLs matched.*

Using these criteria, 127 tools were matched. After uniquifying, the combined list consists of 6092 unique tools; 4832 (79.3%) with at least one associated PMID. Partial investigation suggests that those without PMIDs were majorly database resources. 277 PMIDs which were associated with more than one tool were removed. The remaining PMIDs were used to seed the retrieval of publications citing them. Thus, these PMIDs are referred to as *seeds* and the publications citing them as *citers*. This research used the PubMed cross-citation database maintained by the US National Library of Medicine, National Institutes of Health. The retrievals were done via their Entrez API. Publishing articles of the tools, the *seeds*, are representatives of the tools they publicize. Likewise, articles citing the seeds, the *citers*, are tangible evidences of the tool usages. Thus, descriptors of those articles describes the tools and the usages, respectively. A preliminary assessment has been done by sampling 79 unique citers citing a

single seed to manually evaluate whether a cited tool is actually used; 78 of the 79 citers used the tool. Therefore, the implication was assumed.

To minimize massive missing values, article descriptors which requires processing of full-text articles were excluded. Metadata (in XML format) of the seeds and the citers were retrieved. Figure 3.1 shows a web-based view of the metadata provided by PubMed. Data were extracted from all the XML files. Part of the text processing was done in Python version 2.7. Because PubMed uses Medical Subject Headings (MeSH) to control the vocabularies used in the keywords, the words were decapitalized and used as is. Stemmed words are words in the titles and abstracts of the articles which were stemmed using Port Stemmer algorithm. Functions of the tools were extracted from the function tags in the original tool lists. All the data were stored in a SQL database with the schema shown in Figure 3.2. Data from the database were queried and described using R Language and Environment version 3.3.1 [30]. The major packages used includes RMySQL [31], dplyr [32] and ggplot2 [33].



Mol Genet Genomics. 2014 Aug;289(4):567-73. doi: 10.1007/s00438-014-0831-7. Epub 2014 Mar 9.

OMIGA: Optimized Maker-Based Insect Genome Annotation.

Liu J[1], Xiao H, Huang S, Li F.

Author information

Abstract

Insects are one of the largest classes of animals on Earth and constitute more than half of all living species. The i5k initiative has begun sequencing of more than 5,000 insect genomes, which should greatly help in exploring insect resource and pest control. Insect genome annotation remains challenging because many insects have high levels of heterozygosity. To improve the quality of insect genome annotation, we developed a pipeline, named Optimized Maker-Based Insect Genome Annotation (OMIGA), to predict protein-coding genes from insect genomes. We first mapped RNA-Seq reads to genomic scaffolds to determine transcribed regions using Bowtie, and the putative transcripts were assembled using Cufflink. We then selected highly reliable transcripts with intact coding sequences to train de novo gene prediction software, including Augustus. The re-trained software was used to predict genes from insect genomes. Exonerate was used to refine gene structure and to determine near exact exon/intron boundary in the genome. Finally, we used the software Maker to integrate data from RNA-Seq, de novo gene prediction, and protein alignment to produce an official gene set. The OMIGA pipeline was used to annotate the draft genome of an important insect pest, Chilo suppressalis, yielding 12,548 genes. Different strategies were compared, which demonstrated that OMIGA had the best performance. In summary, we present a comprehensive pipeline for identifying genes in insect genomes that can be widely used to improve the annotation quality in insects. OMIGA is provided at http://ento.njau.edu.cn/omiga.html .

PMID: 24609470 DOI: 10.1007/s00438-014-0831-7

[PubMed - indexed for MEDLINE]

Publication Types, MeSH Terms, Substances, Secondary Source ID

Publication Types

Research Support, Non-U.S. Gov't

MeSH Terms

Amino Acid Sequence
Animals
Base Sequence
Butterflies/genetics*
Exons
Genetic Markers/genetics
Genome, Insect/genetics*
Genomics*
High-Throughput Nucleotide Sequencing
Introns
Molecular Sequence Annotation*
Molecular Sequence Data
Moths/genetics*
Open Reading Frames
Oryza/parasitology
Plant Diseases/parasitology
Repetitive Sequences, Nucleic Acid/genetics
Sequence Alignment

Figure 3.1: A web screenshot of an article's metadata provided by PubMed

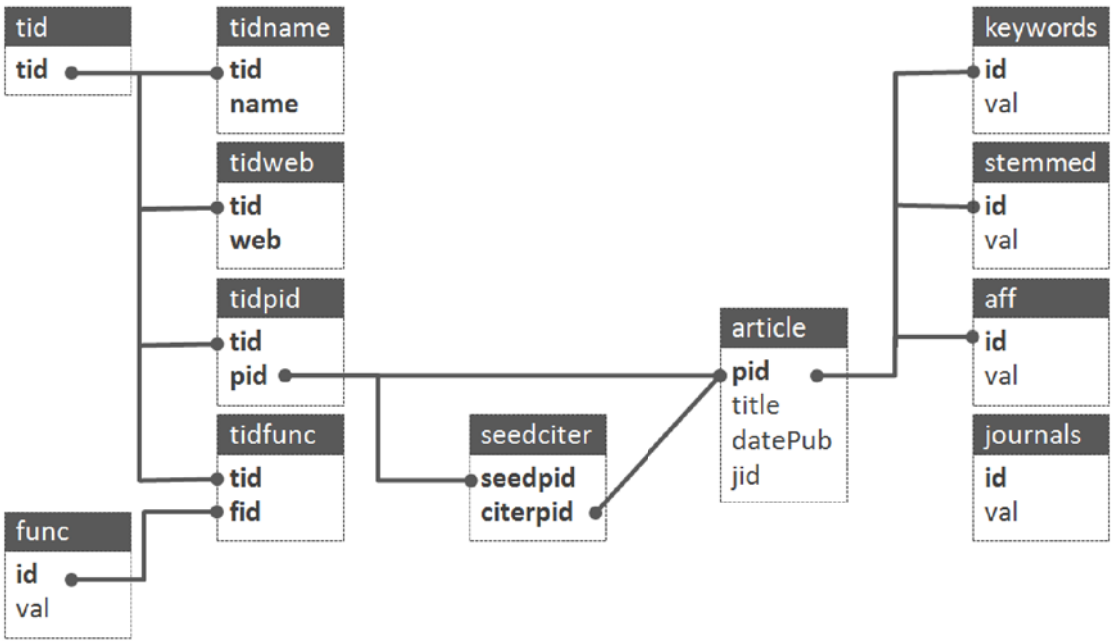


Figure 3.2: SQL Schema to store extracted data

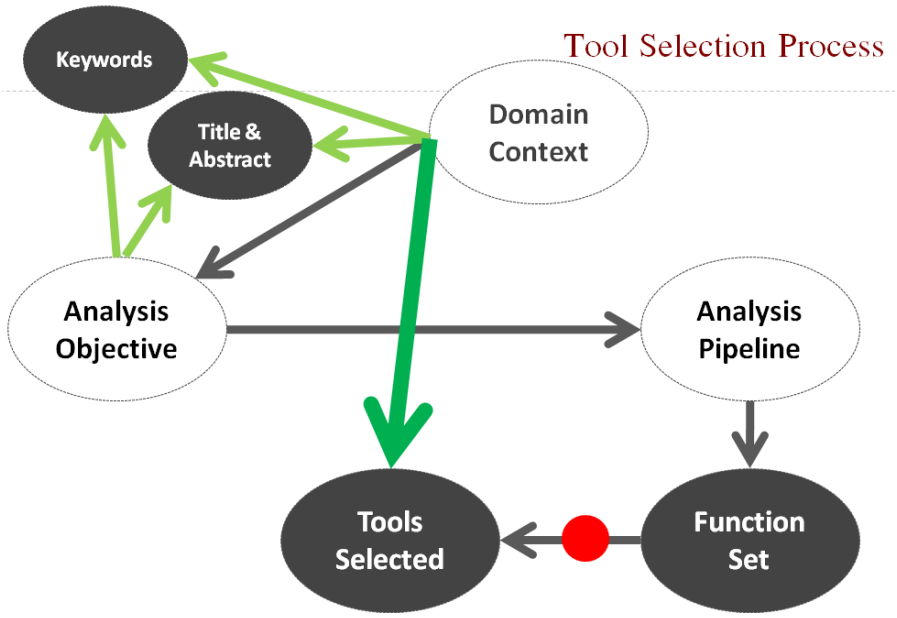


Figure 3.3: The assumed tool selection process: the relationships between the observed and the latent variables

3.2 Constructing the recommendation models

3.2.1 The assumed tool selection process

Based on the literature and common practices, generally, context in a domain drives the analysis objectives. To achieve the objectives, researchers would design analysis pipelines which consists of sequential analysis tasks required to achieve the goals. A function set is a set of functionalities needed to perform the required tasks. The actual tools used for those functionalities are then chosen. Figure 3.3 diagrammatically illustrates the relationships between the tool selection influencers described above. Considering the retrieved data, the hollow nodes represent the latent variables and the filled nodes represent the observed variables. Directions of the arrows are the believed causal relationships. The green arrow represents the influence of domain context on tool selection when adjusted for the differences in the analysis pipelines.

From the tool selection process, the two direct influencers of tool selection are the function set and the domain context. In bioinformatics, each tool typically provides a particular functionality. Therefore, the set of functions used were assumed as all the functions tagged to the tools in which the citer cited. The domain contexts were indicated through two observable variables, the citer's keywords and the stemmed words.

3.2.2 Assessing the influence of domain context on tool selection

If the assumed tool selection process is correct, citers which used similar function sets and were alike in domain context would use a similar set of tool combinations. To test the assumption, the citer instances were clustered into homogeneous function sets through an iterative process (see Section 3.2.3). For each function set, graphs with citers as nodes and number of keywords in common as edges were constructed and partitioned. Being able to subpartition the function set by their keywords infers that several domain contexts are found to be using similar analysis funciton sets. If the tools used among these domain contexts differ, it can be concluded that domain context does influence tool selection.

3.2.3 Deriving the homogeneous function sets

A graph clustering approach was used to group citer instances into homogeneous function sets. The method uses graphs constructed at two-granularities, i.e. citer vertices linked by number of functionalities in common and citer vertices linked by number of tools in common. The solid black links in Figure 3.5 illustrates the construction of such links between the usage instances. Infomap [25], a clustering algorithm based on random walk, was used to partition the first graph into groups of citers involved with similar functionalities. Further local clustering each of those partitions based on the second graph grouped the citers using similar tool combinations together. The local clusters were merged back together when they have sufficiently similar function presence signatures and were used to replace the initial function sets. The partitioning is considered converged when all local clusters are completely merged as one. The resulting function sets are therefore considered homogeneous. Figure 3.4 illustrates the described process as a diagram. The subsections below provides the details of each step in the method. [1]

3.2.3.1 Partition the citer instances into initial function sets

The citer instances were partitioned into a set of clusters K based on the set of tool functions they were involved with. With the citer instances as vertices, the edge weight between any pair of vertex is defined as the sum of common function occurrences. The solid black links in Figure 3.5 illustrates the construction of such links between the citer instances. After partitioning, each resulting function set K_m is a cluster of citer instances with similar function patterns. This initiation step broadly separates the citer instances into disjoint groups with low functionality overlaps.

3.2.3.2 Partition the citer instances into tool combinations

In this step, the distinct tool pipelines from citer instances performing similar analysis tasks were recovered. Specifically, the citer instances in each function set were further partitioned into a set of tool combinations L_m based on the tools they used. The construction of tool level graphs is illustrated in Figure 3.5 as dotted grey links. $L_{m,n}$ refers to the n^{th} tool combination in the m^{th} function set. The number of citer instances in each tool combination was counted, denoted as $C(L_{m,n})$. The function presence ratio vector of a tool combination $P^{m,n}$ is a column vector with

[1] Portions of this chapter previously appeared as [34]

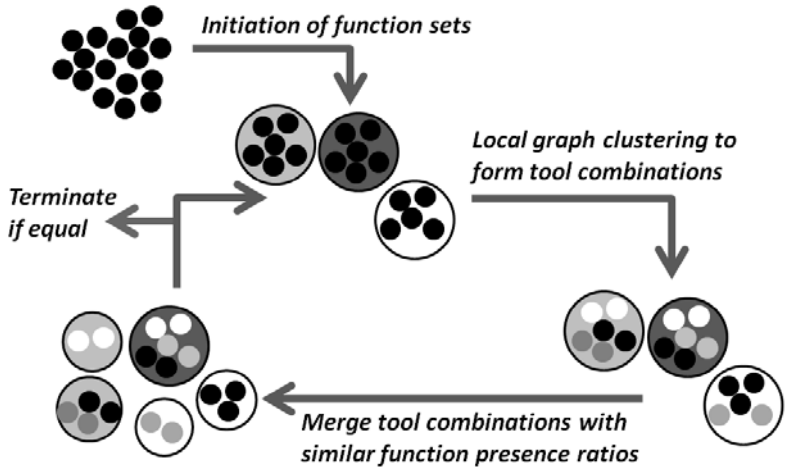


Figure 3.4: A diagrammatic illustration of the iterative graph clustering method

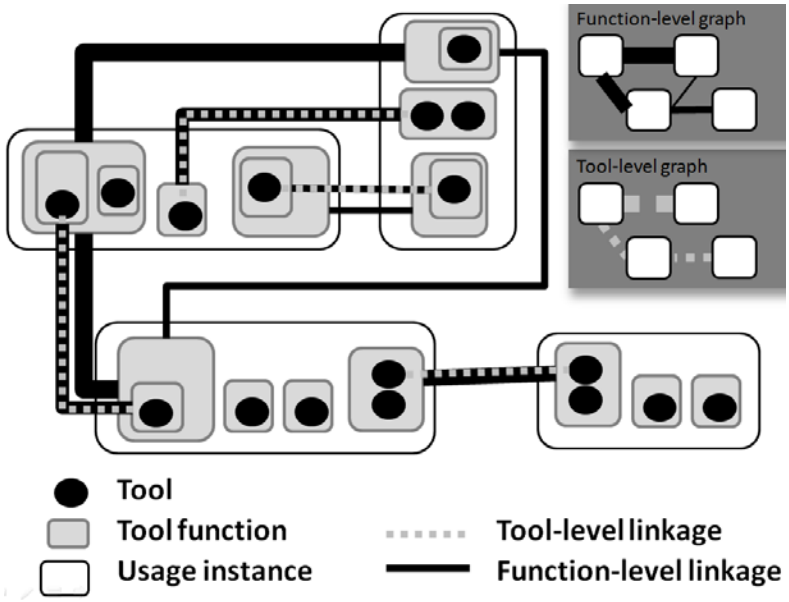


Figure 3.5: An example of the graph topology constructions at the two-granularity layers, i.e. the functionality linkage layer and the tool linkage layer.

each row corresponding to a function f in the tool function space F. The number of occurrence of each function in a tool combination $L_{m,n}$ was counted as $C(F^f_{m,n})$. The values in the vector were calculated according to the Equation 3.1. This vector describes the function distribution of the tool combinations. Ideally, the vector should be identical for all tool combinations within the same function set. In summary, this step groups citer instances using similar tools for the same or similar pipelines together.

$$P^{m,n}_{f,1} = C(F^f_{m,n})/C(L_{m,n}) \tag{3.1}$$

3.2.3.3 Merge the tool combinations

To ensure that the tool combinations homogeneously fit the same function set, we revert the process by merging tool combinations with similar function presence ratios. In this step, each vertex in the graph represents a tool combination. The pairwise similarities between the x^{th} and y^{th} tool combination in the m^{th} function set ($S^m_{x,y}$), calculated from Equation 3.2, are used to represent the edge weights. The merged clusters replace the initial function sets derived from Step 3.2.3.1. A uniform function set results in a complete graph with equal edge weights, and hence, returns a single cluster upon clustering. The tool combination partitioning step and the merging step are repeated until the function sets no longer splits. At the end, each function set groups instances which performed similar analysis tasks together. The tool combinations within each of those function sets are a group of unique tool pipelines supported by usage evidences.

$$S^m_{x,y} = P^{m,x} \cdot P^{m,y} \tag{3.2}$$

where $m \subset M, x \subset N, y \subset N, x \neq y$

3.2.4 The recommendation models

The assumed tool selection process was used to guide the construction of the recommendation models. Each of the competing models included different parts of the variables in the process. Table 3.2 enlists all the models studied in this research.

The table provides the log-likelihood equations used in each model to calculate the tool rankings. Explanations of those equations are also given in the table. The notations used in the equations are defined in Table 3.1. In addition, two approaches of time lag models were explored. The first, Equation 3.3, accounts for the decline in influence of the data over time. The second, Equation 3.4, models the effect of the trend direction (the increase/decline of the loglikelihood over the years) on the log-likelihood. T is the number of retrospective years used in the model. $T = 5$ was used in this study.

$$\text{loglikelihood} = \sum_{t=1}^{T} l_t * \lambda^{t-1} \tag{3.3}$$

$$\text{loglikelihood} = l_1 * 2^{\sum_{t=2}^{T} \frac{l_1 - l_t}{t-1} * \lambda^{t-1}} \tag{3.4}$$

Table 3.1: Definitions of the notations in the equations for tool recommendation rankings

Notation	Definition
Q_f	row-vector of query functions with each element corresponding to a function in the function space
Q_k	row-vector of query keywords with each element corresponding to a keyword in the keyword space
Q_w	row-vector of query stemmed words with each element corresponding to a stemmed word in the stemmed word space
F_T	matrix with each row corresponding to a function in the function space and each column corresponding to a tool; values at each column are the normalized function distribution of the tool
T_T	diagonal matrix with the n^{th} diagonal element corresponding to the n^{th} tool's normalized popularity
F_{S_1}	matrix with the functions in the function space as rows and the initial function sets (IFS) as columns; values at each column are the normalized function distribtuion present in the IFS with IDF weighting
F_{S_m}	matrix with the functions in the function space as rows and the converged function sets (CFS) as columns; values at each column are the normalized function distribtuion present in the CFS with IDF weighting
S_{1T}	matrix with the initial function sets (IFS) as rows and the tools as columns; values at each row are the normalized prevalence of the tool presences in the IFS with IDF weighting
S_{mT}	matrix with the converged function sets (CFS) as rows and the tools as columns; values at each row are the normalized prevalence of the tool presences in the CFS with IDF weighting
C_T	matrix with the converged tool combinations (TC) as rows and the tools as columns; values at each column are the normalized prevalence of the tool presences in the TC with IDF weighting
K_C	matrix with the keywords as rows and the converged tool combinations (TC) as columns; values at each column are the normalized prevalence of the keyword presences in the TC with IDF weighting
W_C	matrix with the stemmed words as rows and the converged tool combinations (TC) as columns; values at each column are the normalized prevalence of the stemmed word presences in the TC with IDF weighting

**IDF weighting* is the inverse document weight which increases the importance of rare descriptors and decreases the importance of generic descriptors

Table 3.2: The tool recommendation ranking models, the used to calculate the tool log-likelihoods, and the descriptions of the models

Model	Log-likelihood equation	Description
ft	$ln(Q_f F_T)$	The similarity of the tool's function to the function query
ftqdrt0	$ln(Q_f F_T(T_T)^{1/4})$	ft model weighted by the 4^{th} root of the tool's popularity (number of usages)
ftoctrt0	$ln(Q_f F_T(T_T)^{1/8})$	ft model weighted by the 8^{th} root of the tool's popularity (number of usages)
fthexdecrt0	$ln(Q_f F_T(T_T)^{1/16})$	ft model weighted by the 16^{th} root of the tool's popularity (number of usages)
ifs	$ln(Q_f F_{S_1} S_{1T})$	The prevalence of the tools in each initial function set (IFS) weighted by the similarity of the IFS function distribution to the function query
cfs	$ln(Q_f F_{S_m} S_{mT})$	The prevalence of the tools in each converged function set (CFS) weighted by the similarity of the CFS function distribution to the function query
ctsk	$ln(Q_k K_C C_T)$	The prevalence of the tools in each converged tool combination (TC) weighted by the similarity of the TC keyword distribution to the keyword query
ctss	$ln(Q_w W_C C_T)$	The prevalence of the tools in each converged tool combination (TC) weighted by the similarity of the TC stemmed word distribution to the stemmed word query
cfs_ctsk	$ln(Q_f F_{S_m} S_{mT}) + ln(Q_k K_C C_T)$	Log-likelihood of the converged function set in conjunction with the log-likelihood of the tool combination (cts model)
cfs_ctss	$ln(Q_f F_{S_m} S_{mT}) + ln(Q_w W_C C_T)$	Log-likelihood of the converged function set in conjunction with the log-likelihood of the tool combination ($ctss$ model)
cfs_ctsk_ctss	$ln(Q_f F_{S_m} S_{mT}) + ln(Q_k K_C C_T) + ln(Q_w W_C C_T)$	Log-likelihood of the converged function set in conjunction with the log-likelihood of the tool combination (cts model and $ctss$ model)
ftoctrt0_ifs	$ln(Q_f F_T(T_T)^{1/8}) + ln(Q_f F_{S_1} S_{1T})$	$ftocctrt0$ model used in conjunction with the initial function sets
ftoctrt0_cfs	$ln(Q_f F_T(T_T)^{1/8}) + ln(Q_f F_{S_m} S_{mT})$	$ftocctrt0$ model used in conjunction with the converged function sets

*The notations in the equations are defined in Table 3.1

3.3 Evaluating the recommendation models

To determine the model which best mimics the tool selection behavior of the field, precision rates and recall rates were measured according to the equations below:

$$Recall = \frac{\text{Number of tools recovered}}{\text{Number of actual tools used}} \tag{3.5}$$

$$Precision = \frac{\text{Number of tools recovered}}{\text{Number of tools recommended}} \tag{3.6}$$

In real use, data of the past are used to recommend the future. Therefore, in assessing the models built from data of year T, input queries with known actual tools used were generated from data of year $T + 1$. In models with time lag incorporated, data of year T,$T - 1$,$T - 2$,$T - 3$ and $T - 4$ were used to build the models. The recommendations were made based on top-N recommendations. Specifically, N tools with the highest log-likelihoods were recommended from each model. The performance assessments were done for N from 5 to 30 with an increment of five; 200 bootstraps at each N with 1000 queries per bootstrap. Bootstrapping is used to estimate the samples from the usage instance population. The significance of the performance differences were compared using independent samples t test for selected models.

CHAPTER IV
RESULTS AND DISCUSSION

4.1 Data Explorations

4.1.1 Describe the dataset

Data were retrieved, extracted, cleansed and stored using the methods described in Chapter 3, Section 3.1. In summary, the dataset consists of 4832 unique tools with at least one associated PMID, the unique identifier of an article indexed by PubMed. These articles are referred to as *seeds* of each tool. 5672 seeds were identified. 10% of the tools were associated with over one seed (see Figure 4.1). This implies a high degree of tool discontinuity.

The seeds were published from a total of 290 journals. 60% of the seeds were published in either Bioinformatics (Oxford, England), Nucleic acids research or BMC bioinformatics (Figure 4.2). Figure 4.3 visualizes the publish date of each seed in each journal. The plot shows the activeness of each journal in contributing to the tool resource.

4.1.2 Describe publishing articles of the tools (seeds)

The publish date of a tool was defined as the most far back publish date among the known seeds. The dataset covers tools published since July 1976 to July 2016. Rate of publication closely fits an exponential equation (Figure 4.4). A high discontinuity rate was observed.

The emergence of tool functions was explored by plotting the tool publish date against the function they were tagged with (Figure 4.5). The vertical blue lines links tool functions tags which co-exist in the same tools. The spanning of the black lines since their first appearance to the far most right and the high connectivity of function tag co-existence suggests persistance of fuctions which appeared in the past till date.

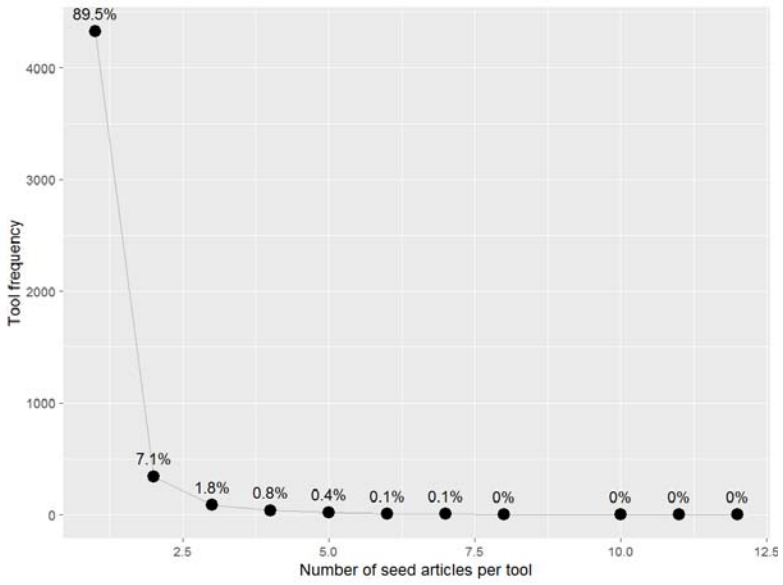


Figure 4.1: The distribution of number of seeds per tool in the dataset

4.1.3 Describe articles citing the seeds (citers)

The 5672 seed articles were queried to PubMed cross citation database via their Entrez API to retrieve PMIDs of all articles citing them, the *citers*. 417433 citers were identified as citing the seeds. 577 seeds have no citations to them. Figure 4.7 shows the number of citations each seed received against the seed's publication date. Though low number of citations are tended towards newer seeds as expected, correlation in number of cites versus time was not observed. In other words, many tools remain subtle despite their long existence.

Of those identified citers, 92466 articles (22.2%) cited only one tool via one seed. 7207 articles (1.7%), though only citing one tool, cited more than one seed. The majority, 317760 articles (76.1%), cited more than one tool. This infers that tools were mostly used in conjunction, i.e. as tool combinations. If the combination consists of highly similar funtionalities, the citer instance is tuitively likely to be a benchmark or review article. Otherwise, if many functionalities were combined, a pipeline may have been used. It should be minded that the one-fourth of citers which cited only one tool may have actually used more than one tool but either the PMIDs they cited were not indexed, or the tools were not included in the two tool lists used in this research.

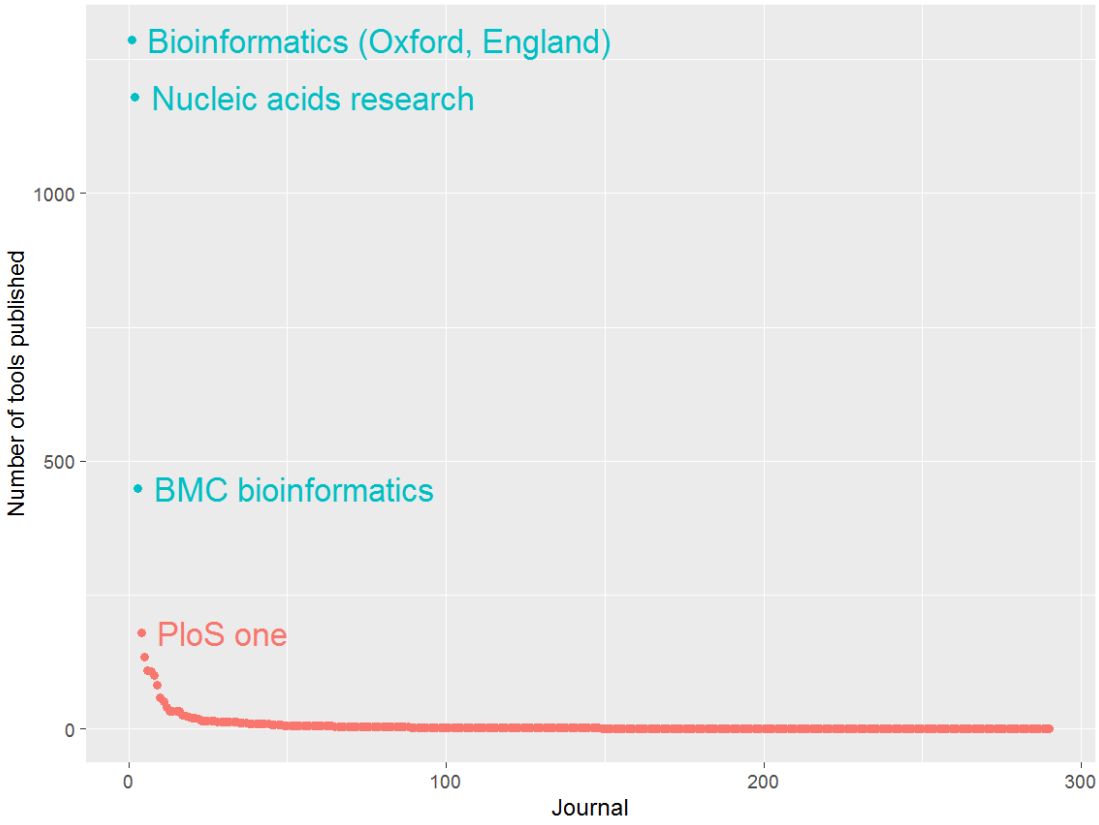


Figure 4.2: The number of seeds published in each journal

Figure 4.7 shows the relationship between the number of tools cited and the number of seeds cited. The diagonal yellow line represents one seed per tool relationship. The majority of citers did not cite several seeds per tool. The citing behaviors which were inconsistent with the mainstream may imply differing citation semantics.

Figure 4.8 shows the number of keywords tagged to the citers during 1988-2016 as a boxplot. Only a few of the citers in 2016 have their MeSH keywords tagged. The availability of the keyword tags in 2015 were low compared the preceding years. Based on the available tags, the emergence and activeness of the keywords over time were examined in Figure 4.9. Waning of the emerged keywords was not apparent. Yearly keyword prevalence was calculated as the proportion of citers containing a keyword to the total citers identified in the year. The prevalence is visualized in Figure 4.10. From the plot, the highly prevalent keywords were similar accross the years. These words appeared to be common concepts in bioinformatics.

4.2 Recommendation model constructions and evaluations

4.2.1 Function set derivations and the domain context

Data of the citers from year 2009-2016 were used to derive yearly function sets as per the methods described in Chapter 3, Section 3.2.3. None of the function sets derived from these years were able to be subpartitioned into clusters with differing keyword attributes. This suggests that the keywords are similar among the citers in each of the function sets. However, it cannot be assumed that the tools used were similar. In the tool combination partitioning step of the function set derivation (as described in Chapter 3, Section 3.2.3.2), a large number of subpartitions were found.

4.2.2 Evaluations of the recommendation models

Recommendation models were built from data of year 2009-2015, separately for each year. Data from the adjacent subsequent year was used as input queries for

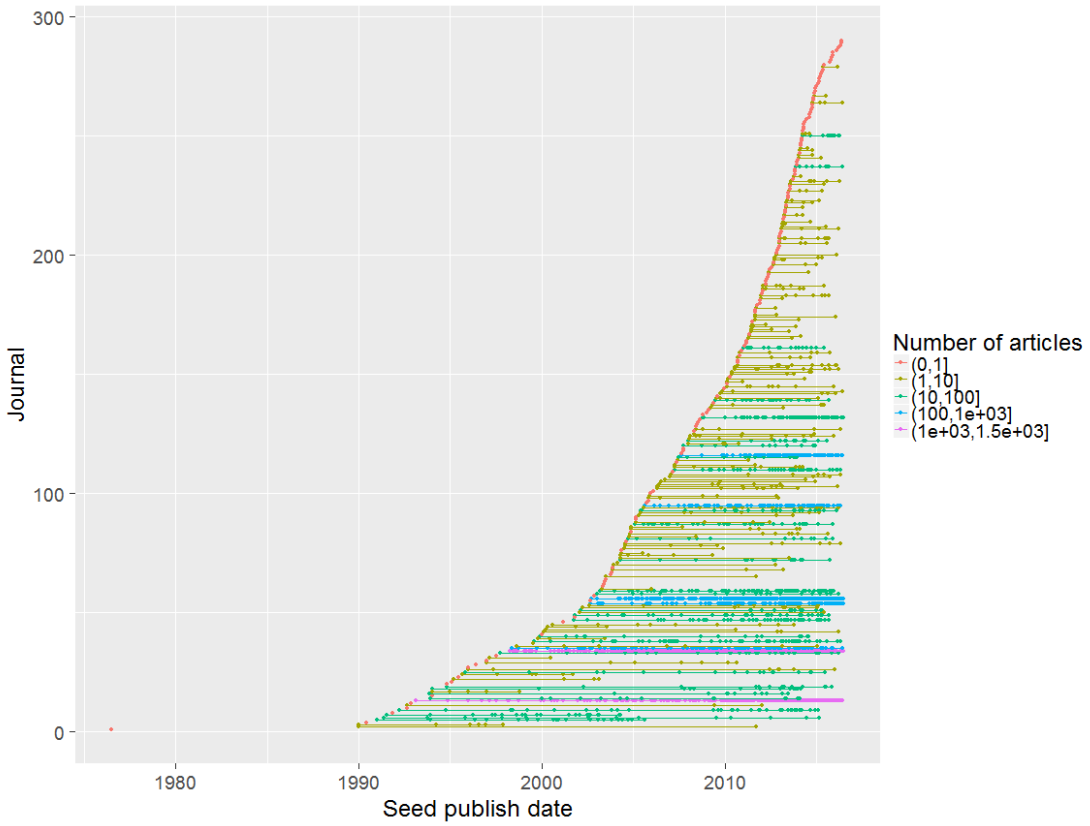


Figure 4.3: The seed publish dates in each journal

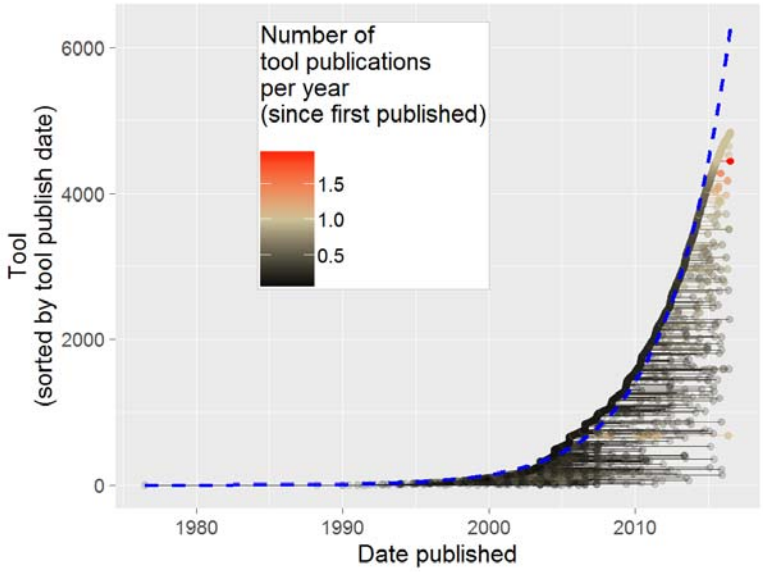


Figure 4.4: Rate of tool emergence overlaid with an exponential fit line (blue dashed)

the evaluations. The mean of the precisions and recalls from the bootstraps at each N of the top-N recommendations were plotted in Figures 4.11-4.17. Overall, the model performance rankings were consistent across all N's. Model *ftoctrt0* appeared as the best model for all the years followed by the *fthexdecrt0* model. Both models relies on matching of query functionalities to the individual tools with their usage popularities taken into account. The achieved recall rates of the two models at N=10 were all above 75% with precisions above 13%. In other words, from the ten tools recommended, 1-2 actual used tools were recovered and 0-1 of the tools used were missed. For one-fourth of the instances, this is the highest achievale precision since they only cited one tool. The ties between the two models were more observed in the more recent years.

The next runner ups were the *ft_cfs*, *ftoctrt0_cfs*, and *ftqdrt0* models. The lesser fit of these models compared to the first two infers that the community considers the tool popularities in general disregard of the analysis pattern being performed. For the queries of 2013-2015, the performance of the three runner ups were comparatively distant from the two best models than the other years. In year 2016, however, the *ftoctrt0_cfs* model had equivalent performances with the two best models. Likewise, the *ft_cfs* model and the *ft* model performed very similarly. This suggests that the derived function sets from the iterative clustering method closely resembles the actual analy-

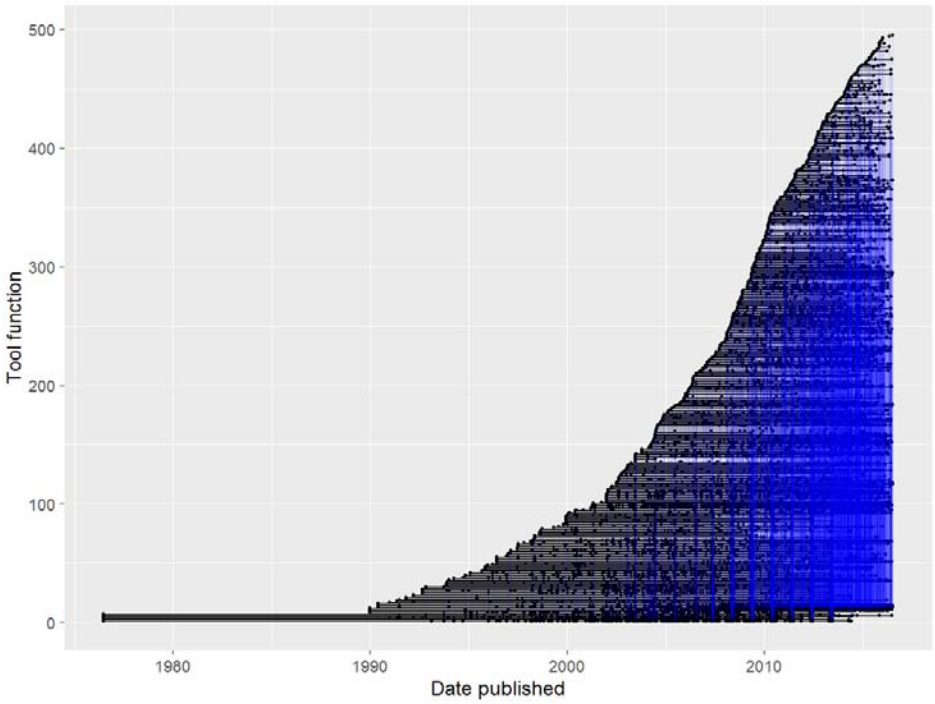


Figure 4.5: Emergence and activeness of the tool functions (black) and their co-occurrence in a tool (blue)

sis patterns. It should be noted that the performance of the *cfs* model was indeed low. An explanation is that the instances were only using a part of the recovered analysis pipeline. And therefore, not offseting the tool rankings within the function sets with the part desired results in a negative impact on the recommendation performance.

The midrange performing models were *ft_ifs*, *ftoctrt0_ifs* and *ifs*, respectively. These models relies on the single step usage instance clustering. The inferior performance of these models to the models utilizing iterative clustering confirms that the iterative clustering method proposed in this research improves the quality of the clusters. That is, the clusters are more homogeneous and better reflects the real patterns in the data.

The domain context involved models (i.e. *ctsk*, *ctss*, *cfs_ctsk* and *cfs_ctss*) performed poorly for all the years. Such result implies that the influence of domain context on tool selection is low. Nonetheless, many tool combinations were identified within the derived function sets. Future studies are needed to understand the underlying factors of such differences.

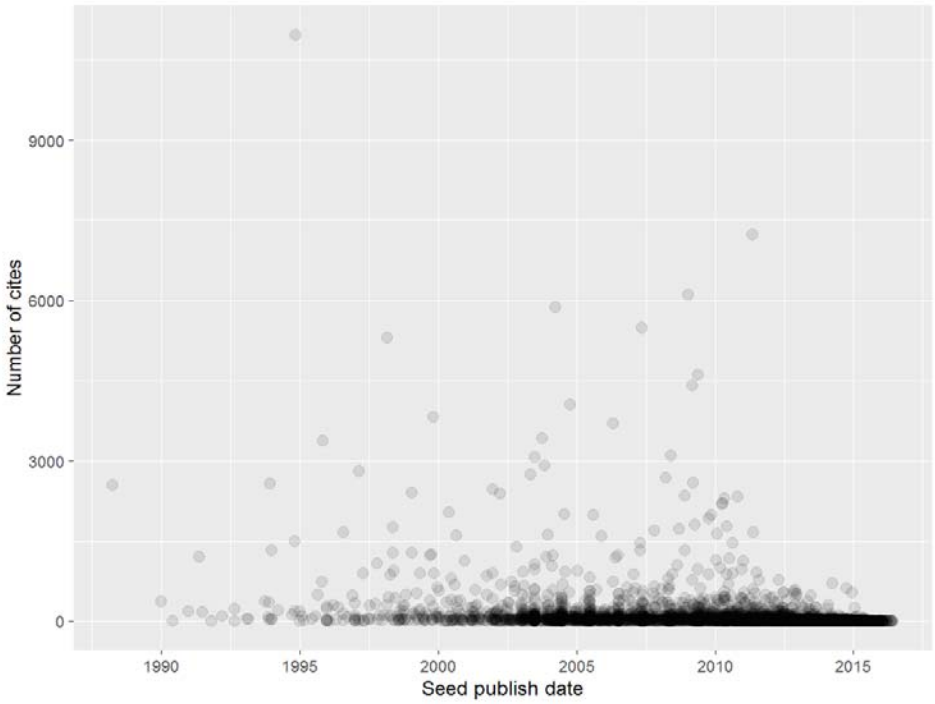


Figure 4.6: The total number of citations a seed received shown with respect to the seed's publish date

The fact that functionalities of the individual tools coupled with their global popularities outperformed models with usage context incorporated (i.e. the combinatory of the functionalities needed and the domain context) is indeed a reflection of the research practice in general. To publish a research, the methodology used must be agreed by the publisher's editor or reviewer. Such constraint biases towards the use of globally popular tools and decreases the likeliness of trying out novel tools. A higher reputation of the author is needed to convince that the novel tools are indeed more suitable in the specific analysis context.

For queries generated from data of year 2014-2016, two approaches of 5-years time lag models were built according to Equations 3.3 and 3.4 defined in Chapter 3, Section 3.2.4. Specifically queries from 2014 were used to assess the time 5-years time lag models built from data of 2009-2013 (Figures 4.18 and 4.19); models from data of 2010-2014 were assessed by queries from 2015 (Figures 4.20 and 4.21) and models from data of 2011-2015 were assessed by queries of 2016 (Figures 4.22 and 4.23). Both time lag models did not improve the best models built from one-year data. To aid the

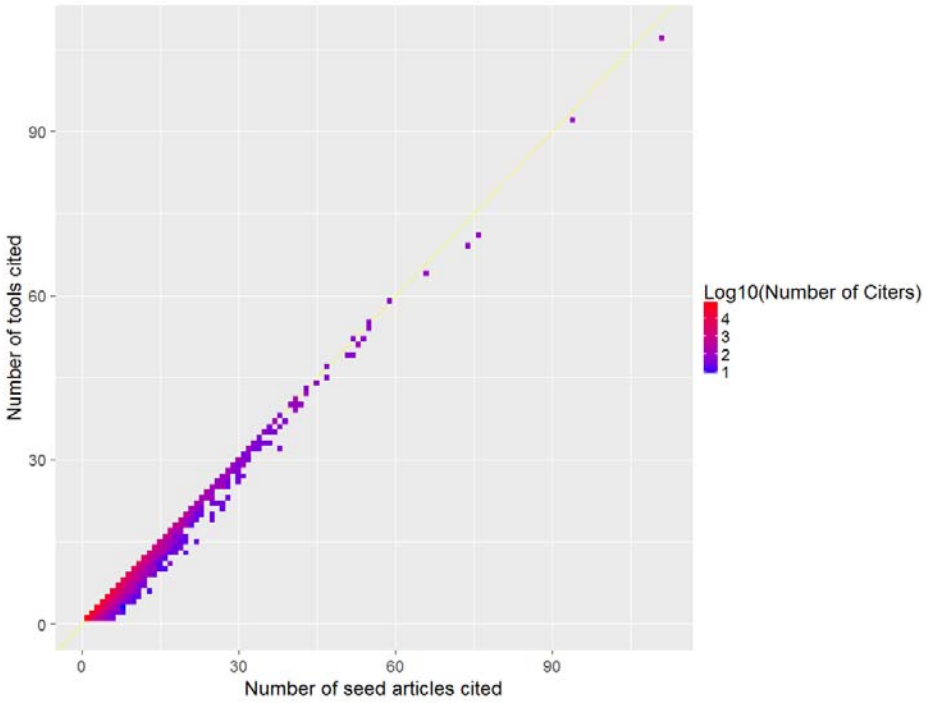


Figure 4.7: A heatmap of seed citing patterns (the number of tools cited versus the number of seeds cited)

interpretations of the time lag models, recommendation models built from two and three years in prior the queries were evaluated as well for these queries; performances shown in Figures 4.24-4.29. The observed performance of models built from data of year $T-1$ to predict the tool choices of year $T+1$ was similar to models from year T. The performance highly dropped when using data of year $T-2$. This suggests that the practices considerable change after two years. An explanation of the performance decline could be the rate of discontinuity observed in Figure 4.4 in conjunction with the growing usage of newer tools used by reputable pioneers. Identifying these early adopters may be beneficial in accelerating the the disuse of discontinued tools, promoting the quality of research results and knowledge derived.

4.3 Tool recommendations for function sets in 2016

Function distributions of the function sets derived from citers of 2016 were queries against the *ftoctrt0* model, the overall best fit model. Figure 4.30 is a screenshot

Figure 4.8: Number of keywords tagged to each citer from 1988-2016

of the recommender user interface (UI) built. The UI provides a slider bar to nagivate through all the function sets derived from data of year 2016. Once a function set is selected, the top table in the page describes the function set through the top-5 functions associated with it. Two function rankings were used, the term frequency (TF) and the term frequency - inverse document frequency (TF-IDF). The table underneath gives a list of top recommended tools for the functions associated with the function sets. The plot on the right, the evidence clustering plot, shows the number of evidences supporting each function set in a log10 scale. The subpartitioning of the function sets over the iterations were also visualized to inform the users of the relationships among the function sets. A highlight bar is overlaid on the plot to accent the function set which the user is viewing. The interface is designed to provide a summarized view of the occurring bioinformatics analysis patterns at a point in time. Experts in the field can review the information to keep up-to-date with the status-quo. For newcomers of the field or users who are reaching out for new analysis pipelines, the information provides an alternative

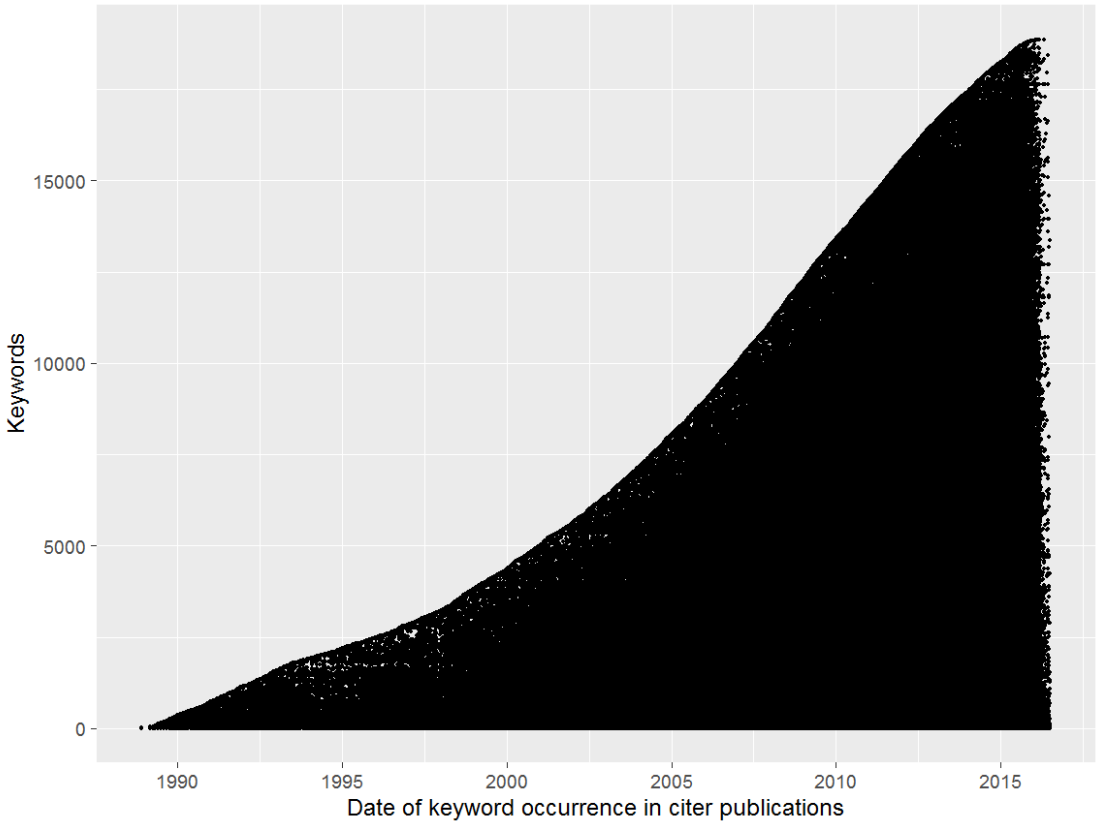


Figure 4.9: The emergence of keywords in citers over time

to the approaches used today. Instead of tracing the literatures to identify the probable candidate tools, the interface provides a targetted entry point to the relevant resources to shorten their literature gathering and reviewing time.

Figure 4.10: The emergence of keywords in citers over time

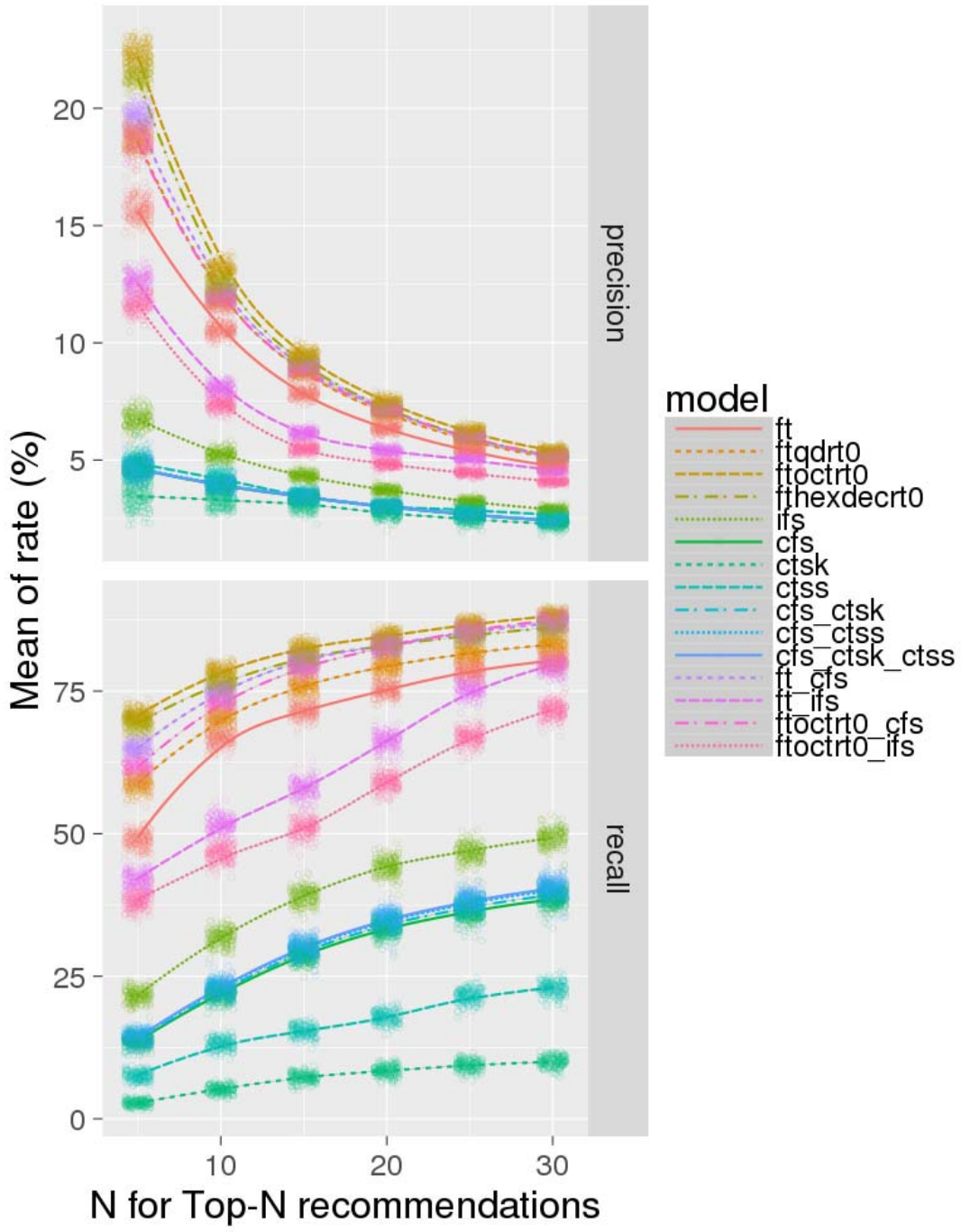


Figure 4.11: The precisions and recalls of the top-N recommendations for models built from data of year 2009 queried by data of year 2010

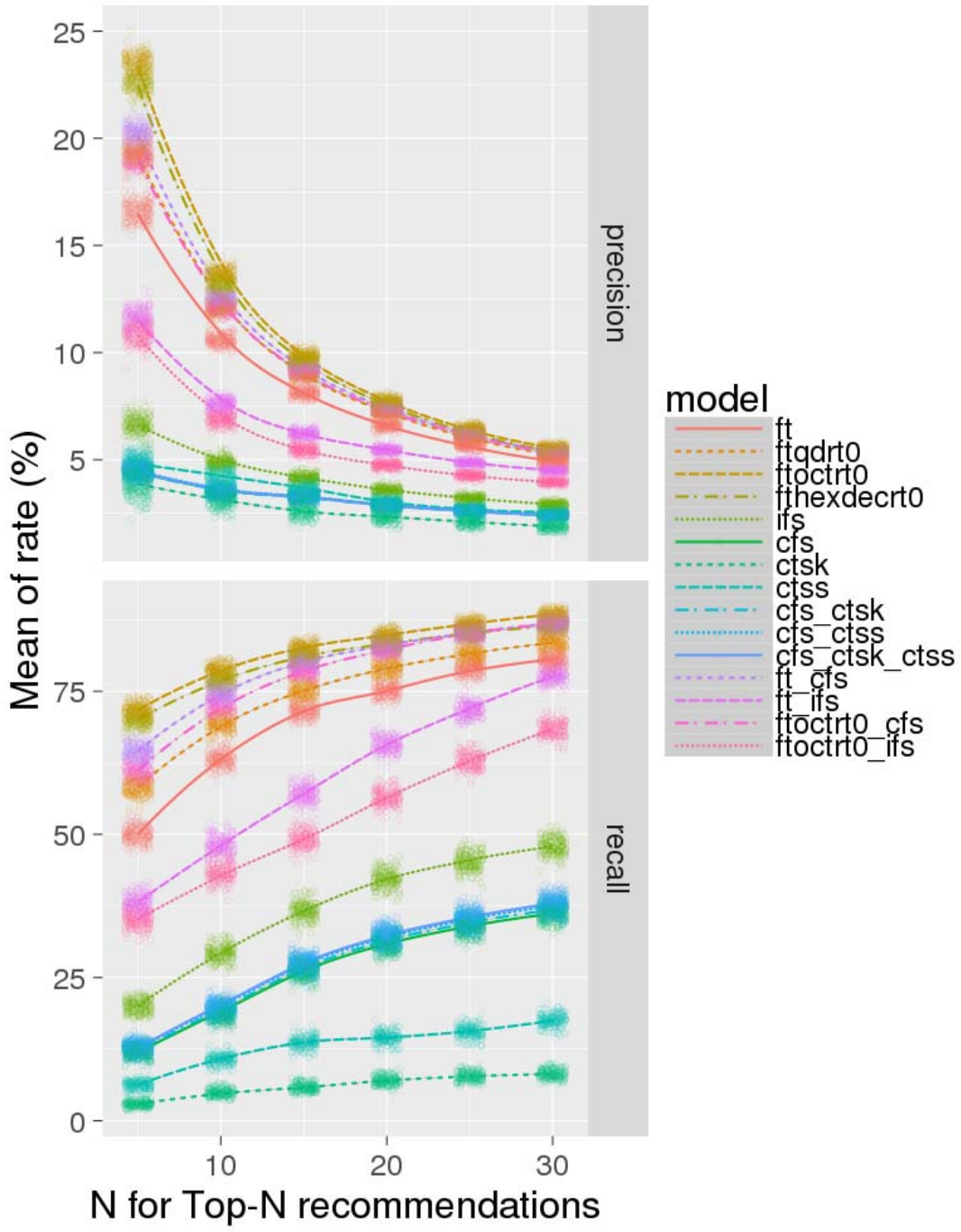


Figure 4.12: The precisions and recalls of the top-N recommendations for models built from data of year 2010 queried by data of year 2011

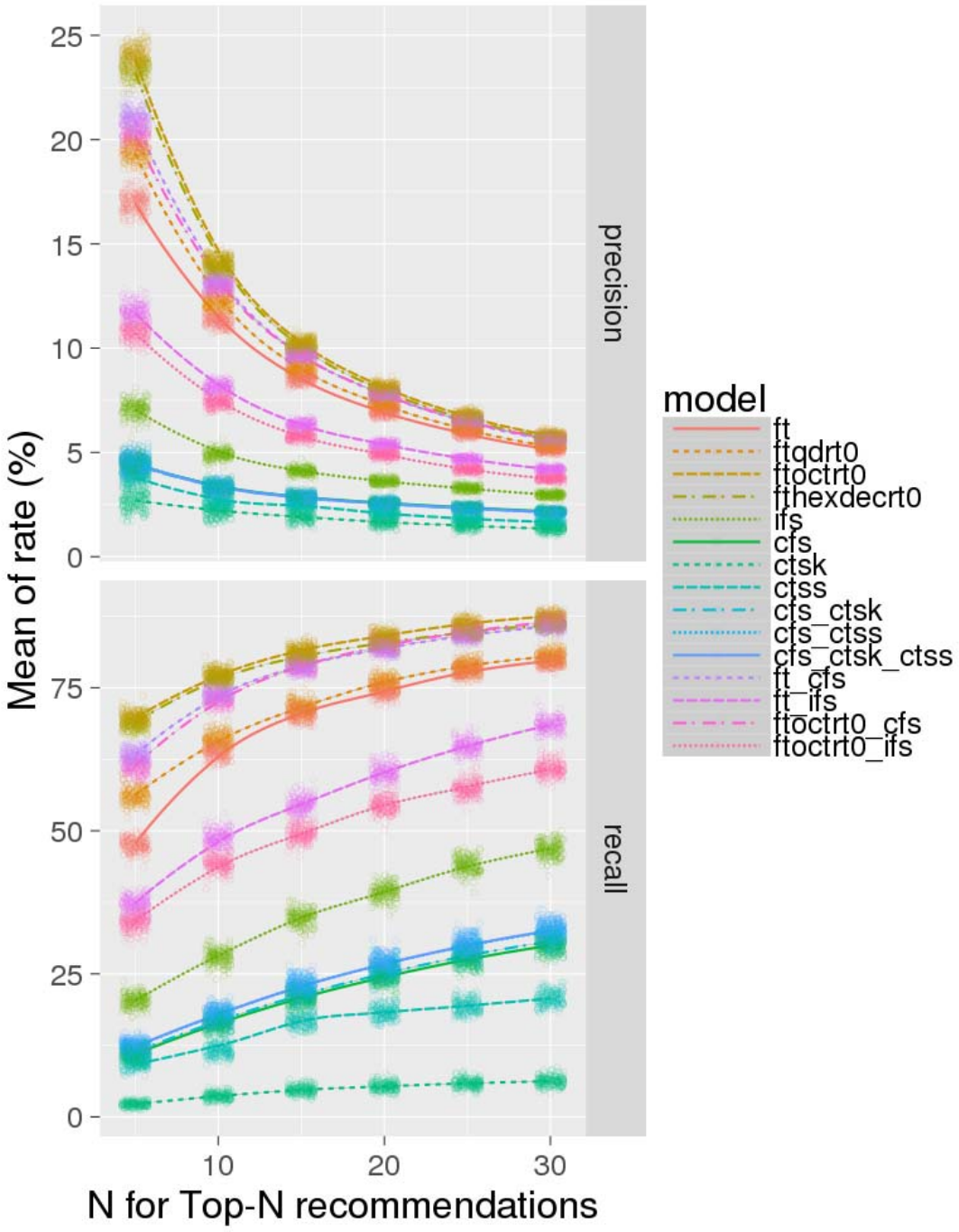


Figure 4.13: The precisions and recalls of the top-N recommendations for models built from data of year 2011 queried by data of year 2012

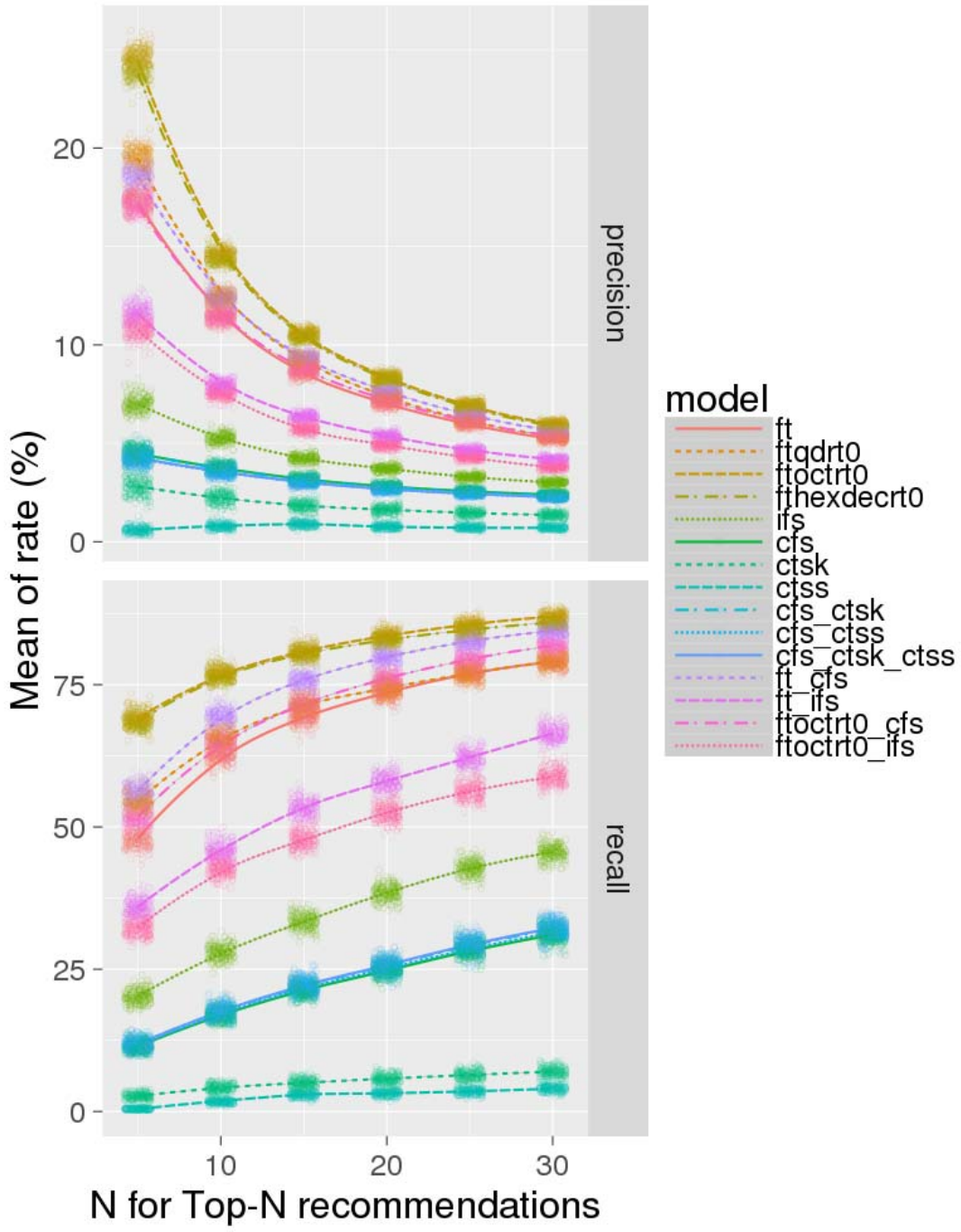


Figure 4.14: The precisions and recalls of the top-N recommendations for models built from data of year 2012 queried by data of year 2013

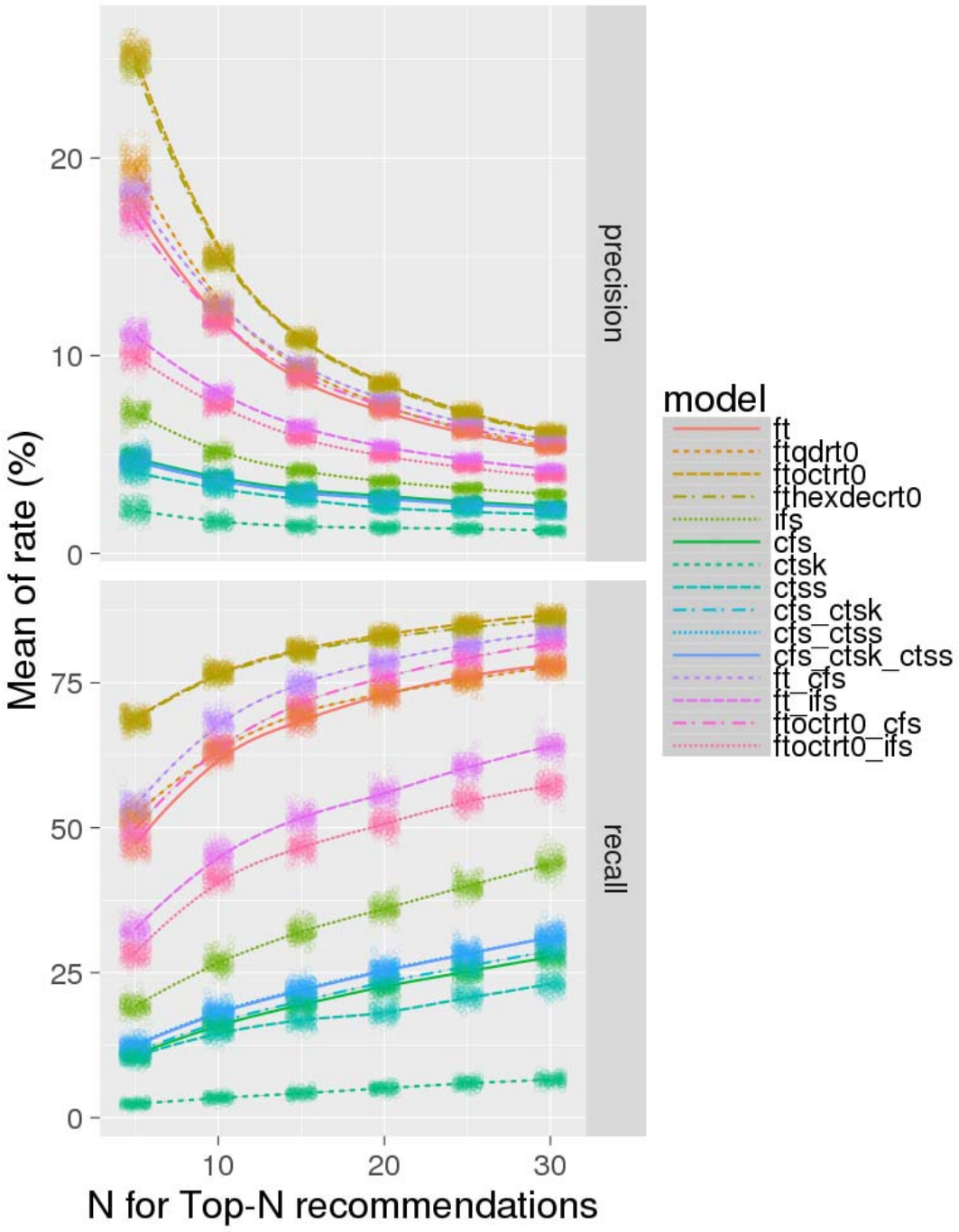


Figure 4.15: The precisions and recalls of the top-N recommendations for models built from data of year 2013 queried by data of year 2014

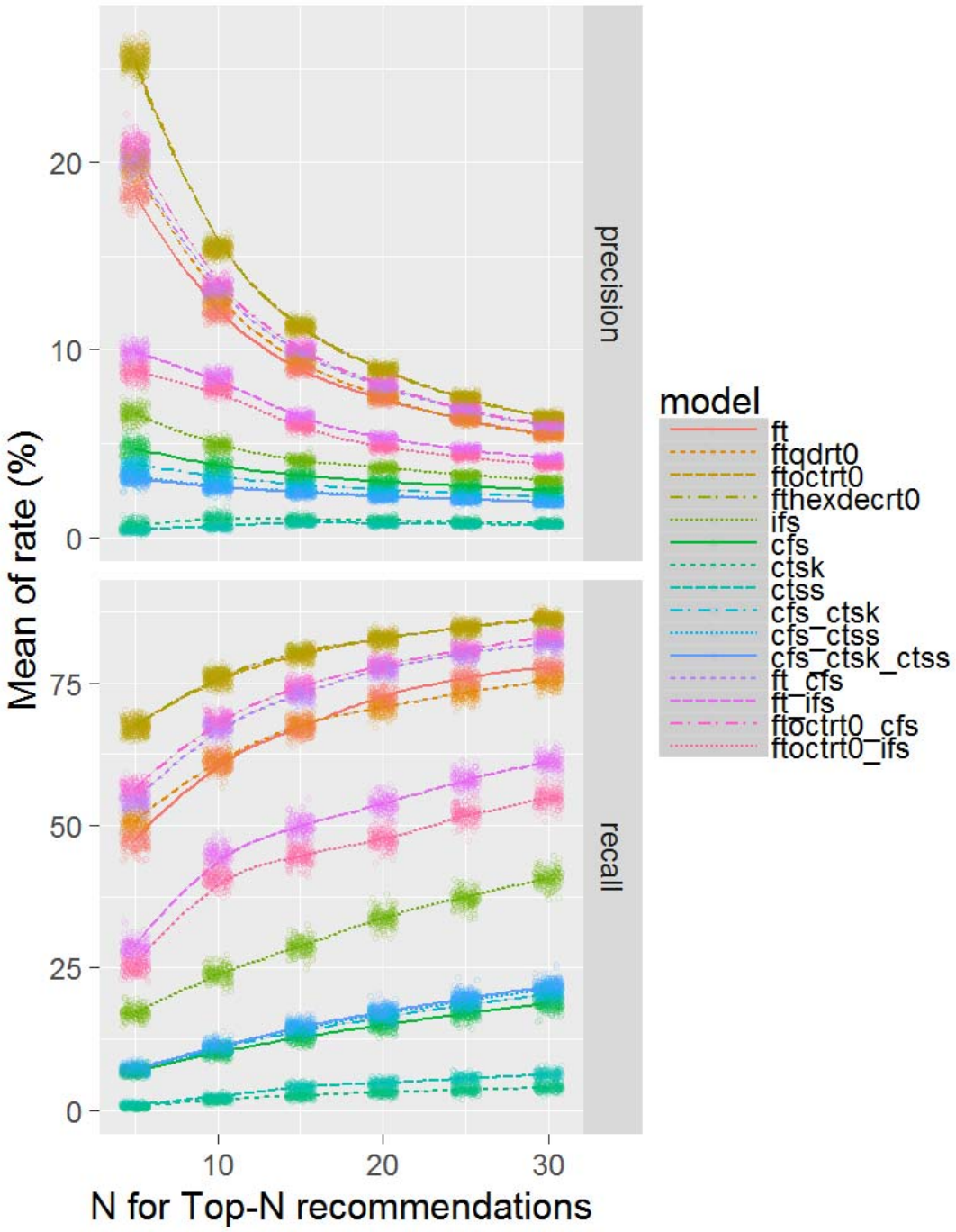


Figure 4.16: The precisions and recalls of the top-N recommendations for models built from data of year 2014 queried by data of year 2015

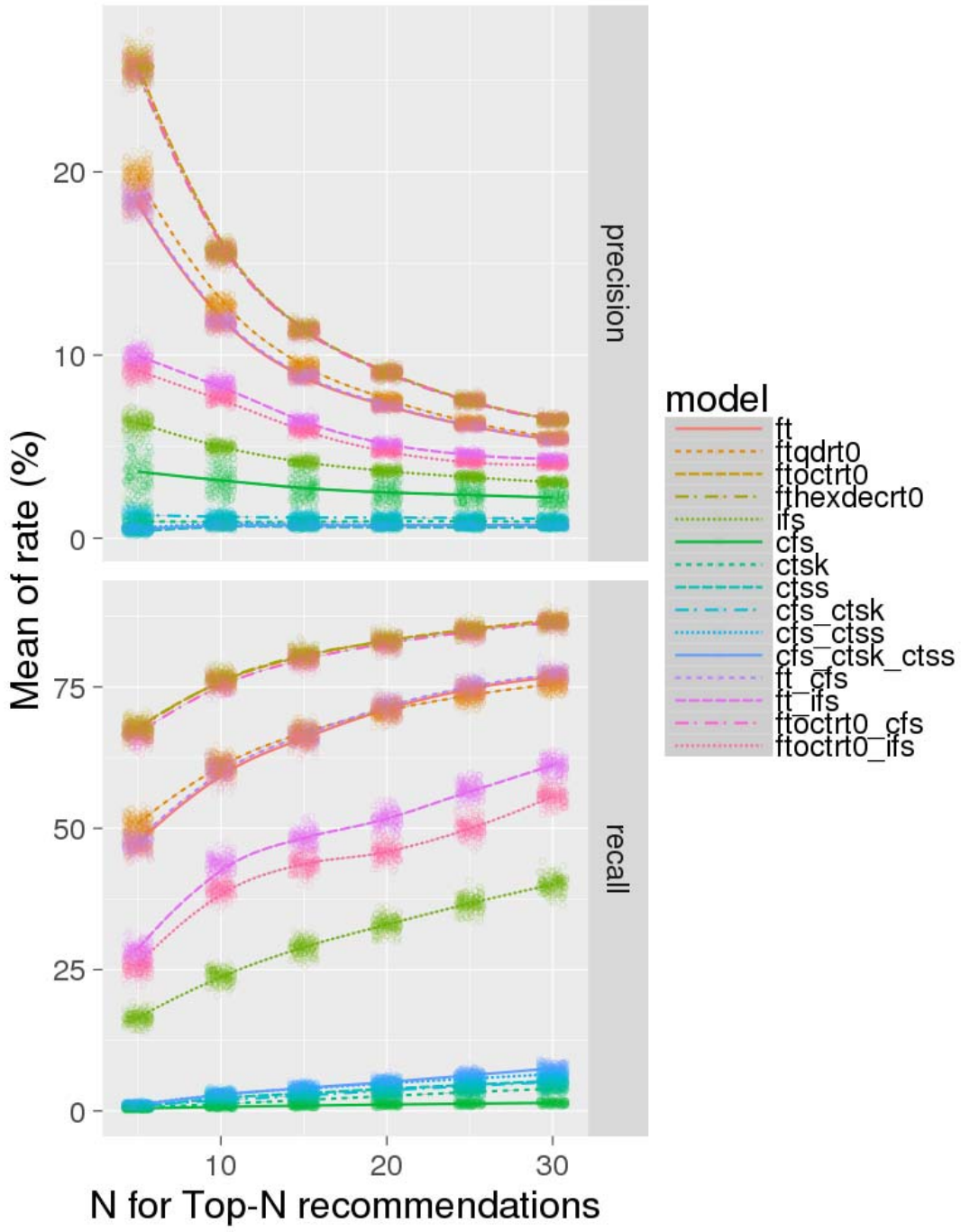


Figure 4.17: The precisions and recalls of the top-N recommendations for models built from data of year 2015 queried by data of year 2016

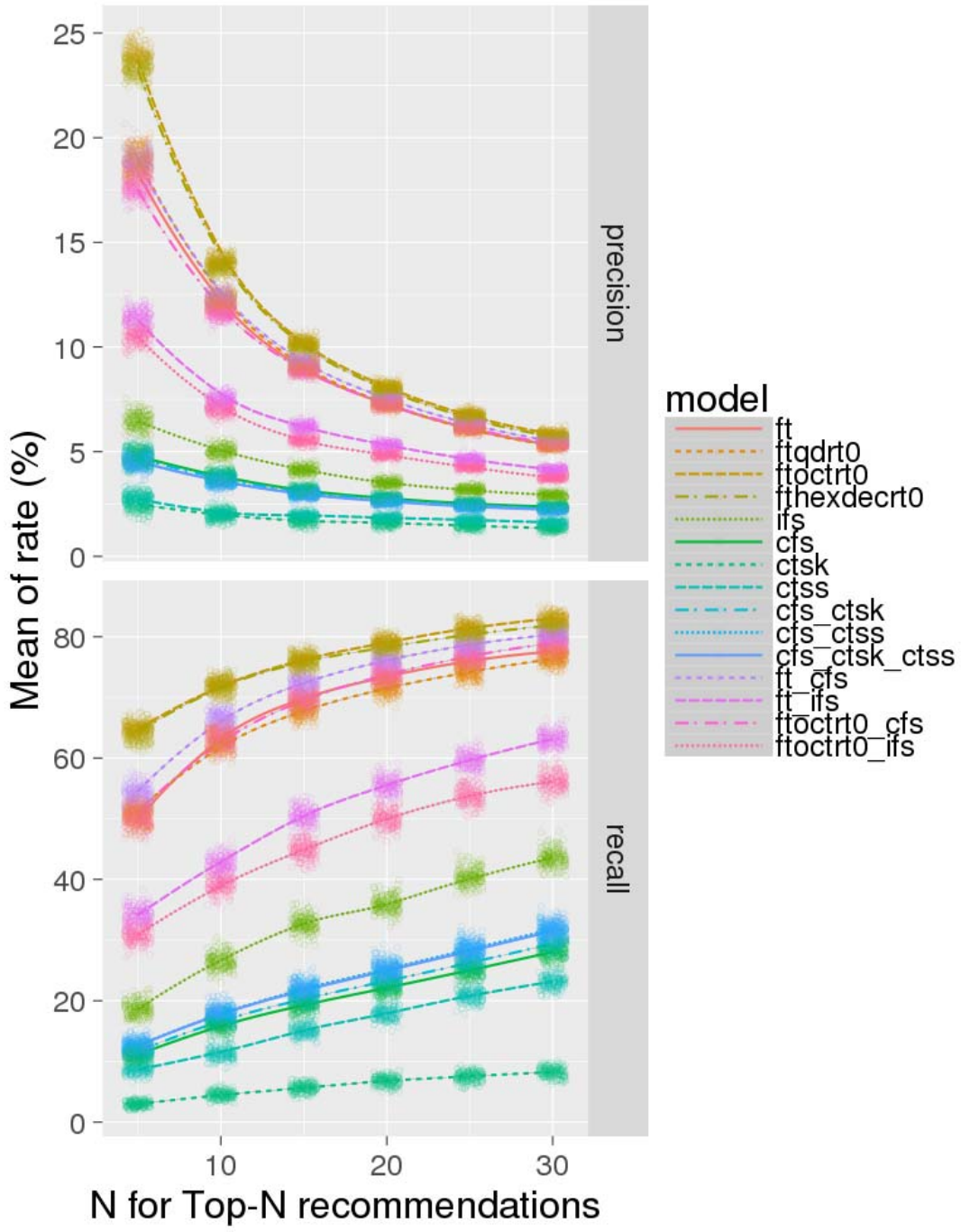


Figure 4.18: The precisions and recalls of the top-N recommendations for time lag models built from data of year 2009-2013 (with declining influence) queried by data of year 2014

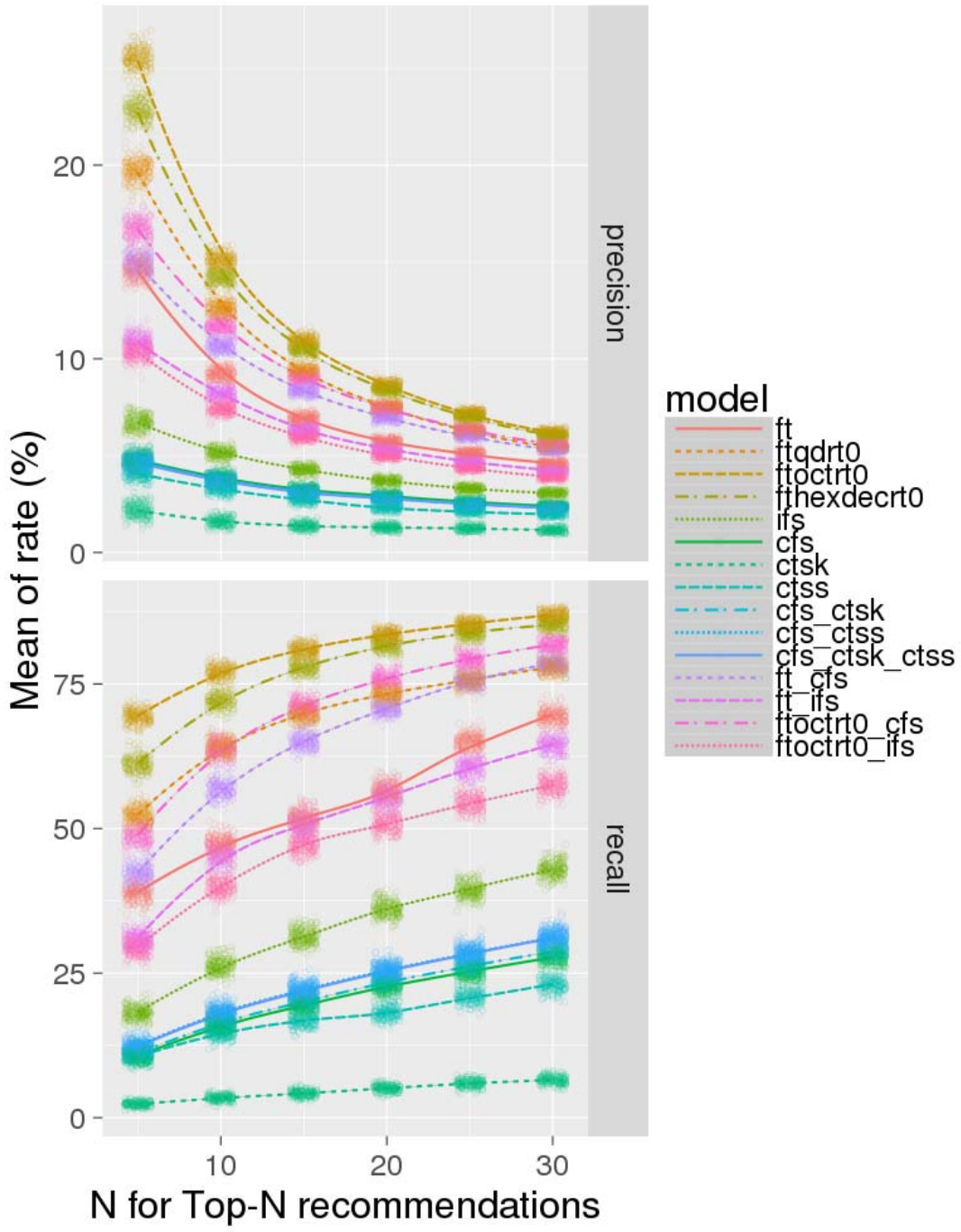


Figure 4.19: The precisions and recalls of the top-N recommendations for time lag models built from data of year 2009-2013 (with trend adjustment) queried by data of year 2014

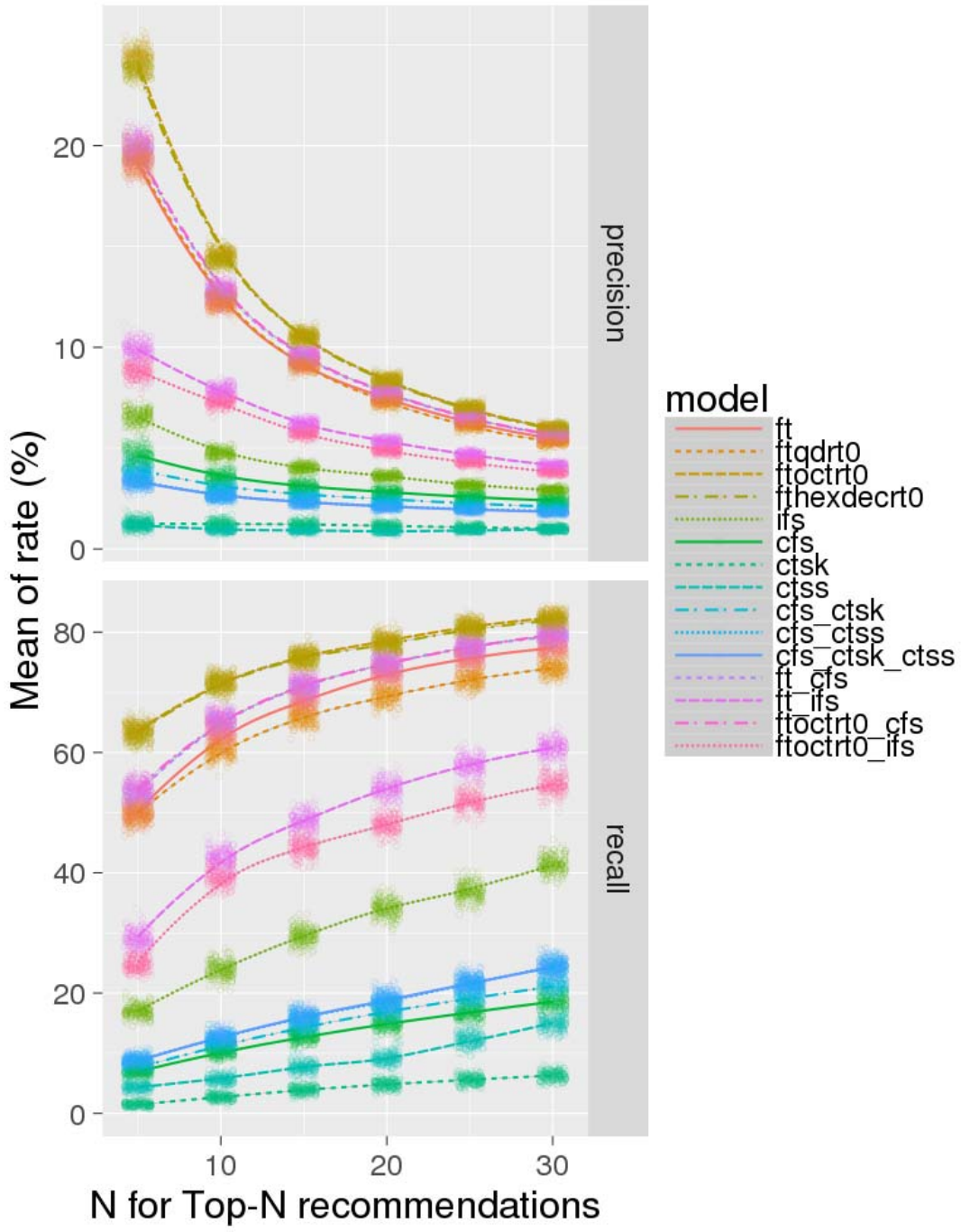


Figure 4.20: The precisions and recalls of the top-N recommendations for time lag models built from data of year 2010-2014 (with declining influence) queried by data of year 2015

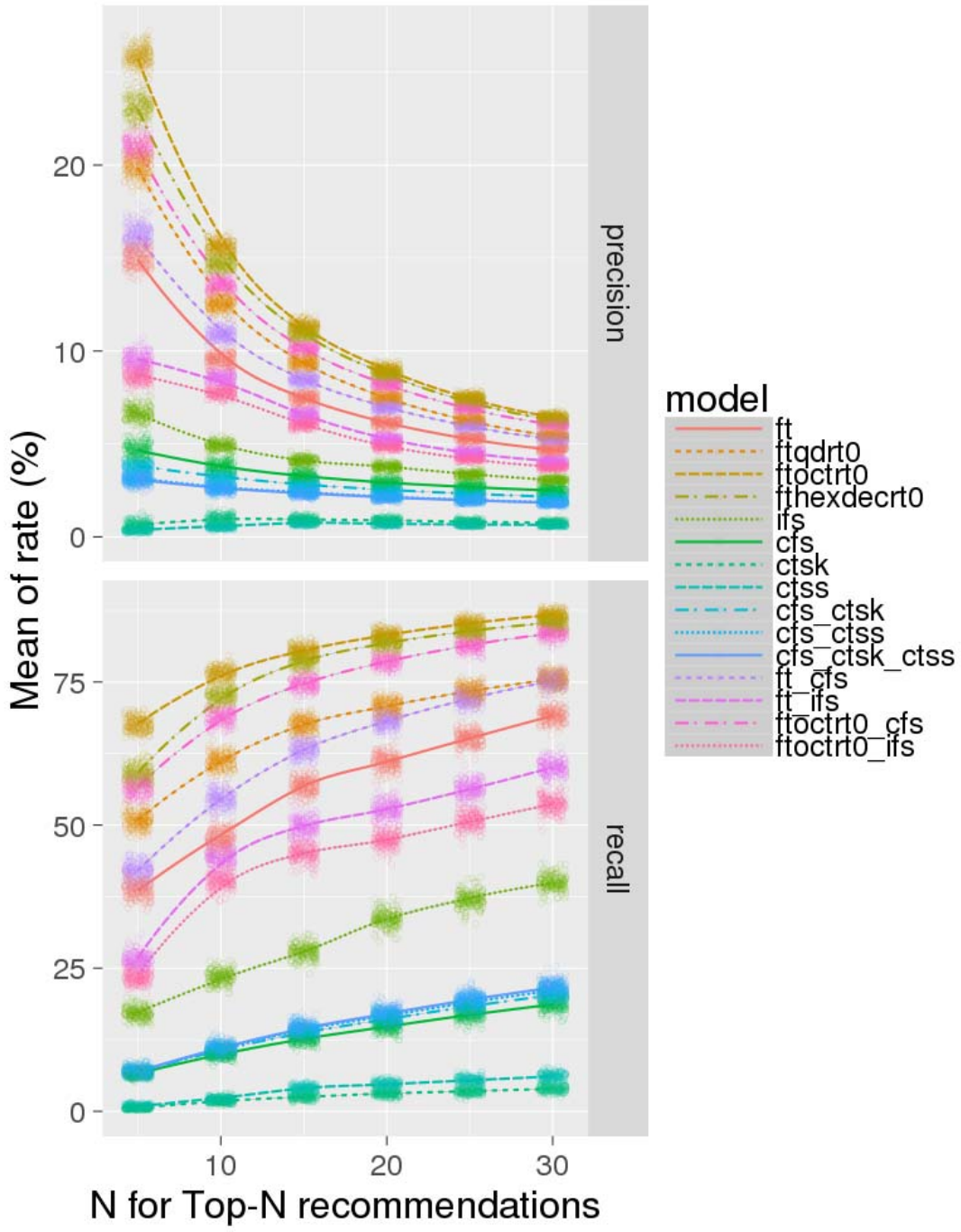


Figure 4.21: The precisions and recalls of the top-N recommendations for time lag models built from data of year 2010-2014 (with trend adjustment) queried by data of year 2015

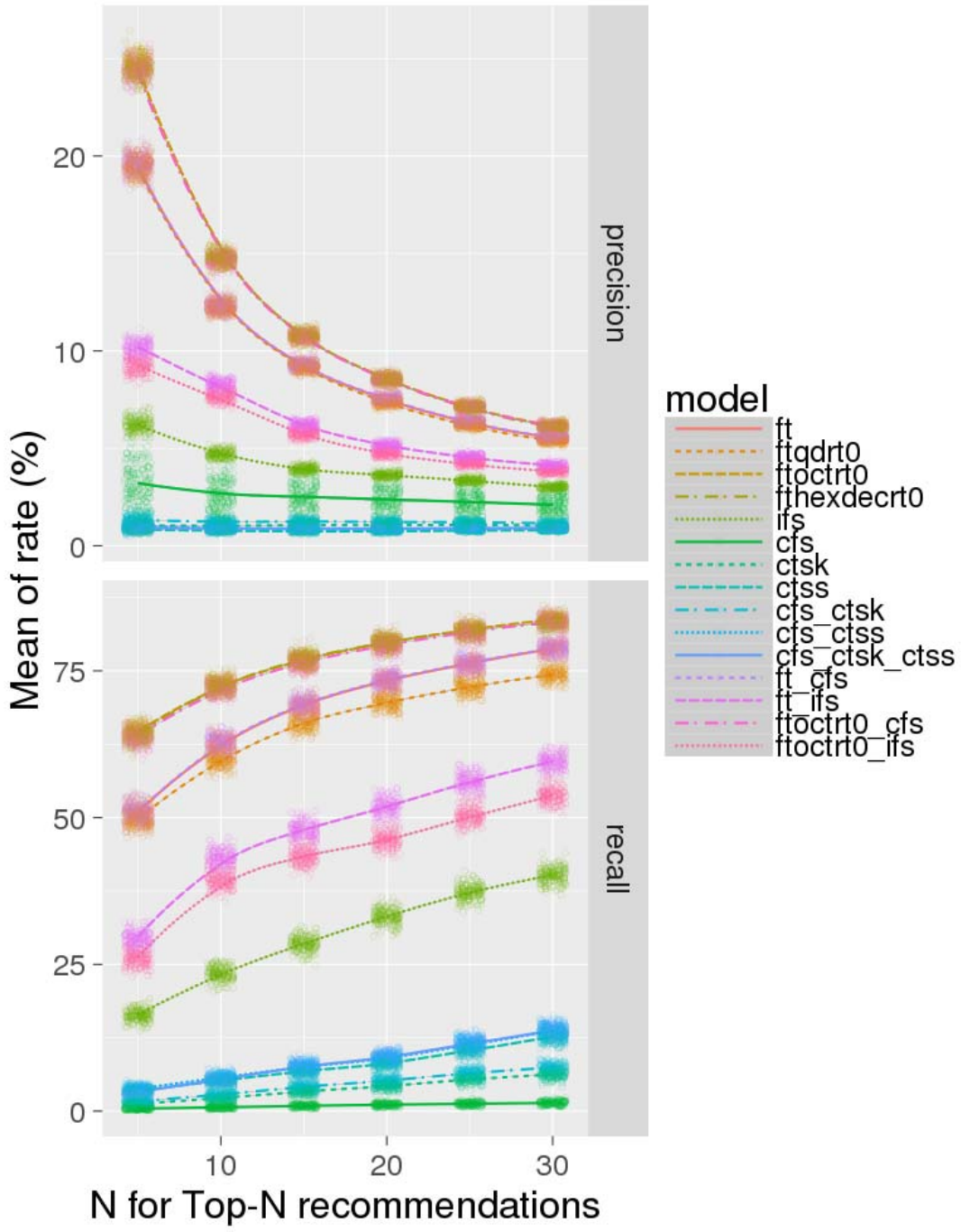


Figure 4.22: The precisions and recalls of the top-N recommendations for time lag models built from data of year 2011-2015 (with declining influence) queried by data of year 2016

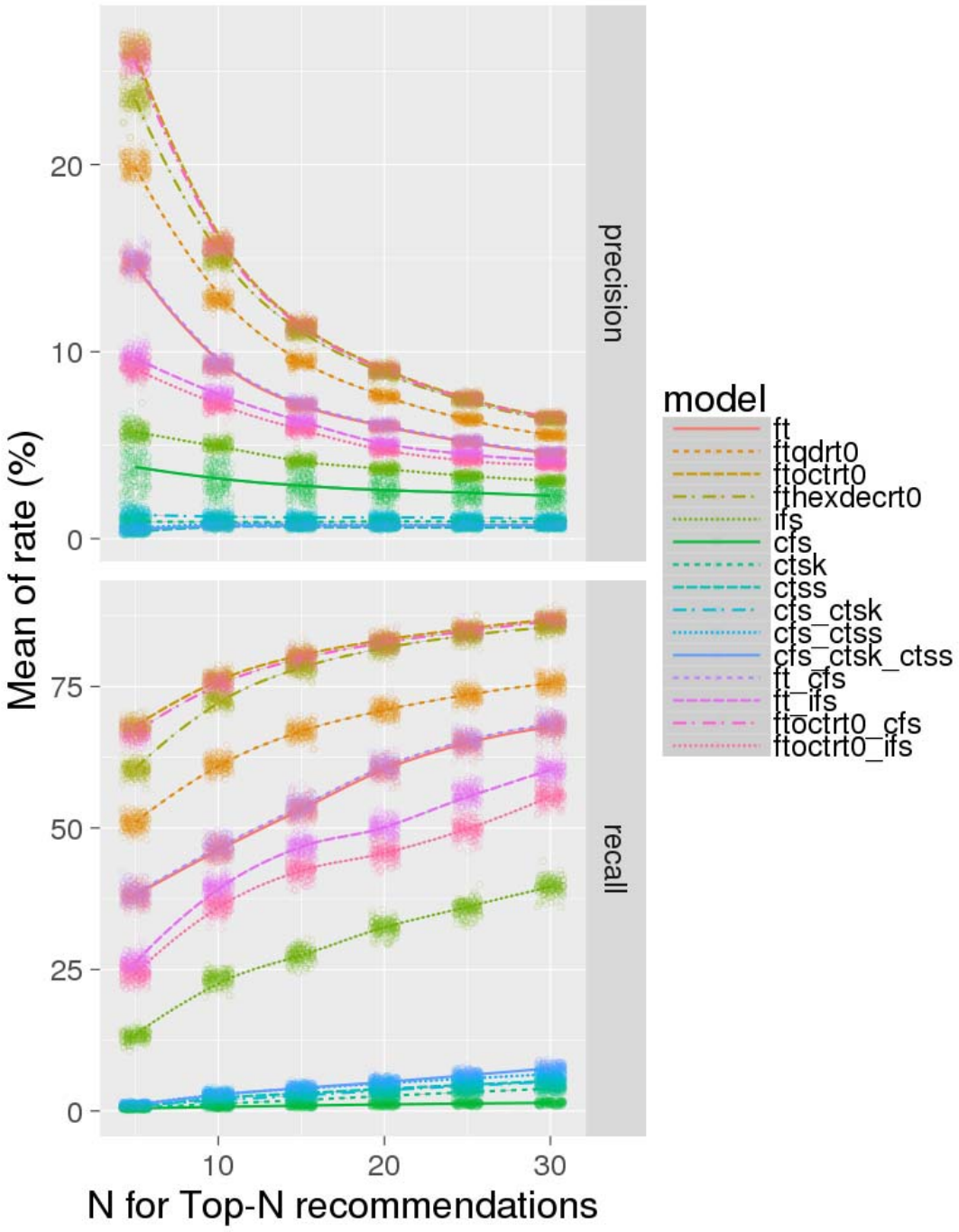


Figure 4.23: The precisions and recalls of the top-N recommendations for time lag models built from data of year 2011-2015 (with trend adjustment) queried by data of year 2016

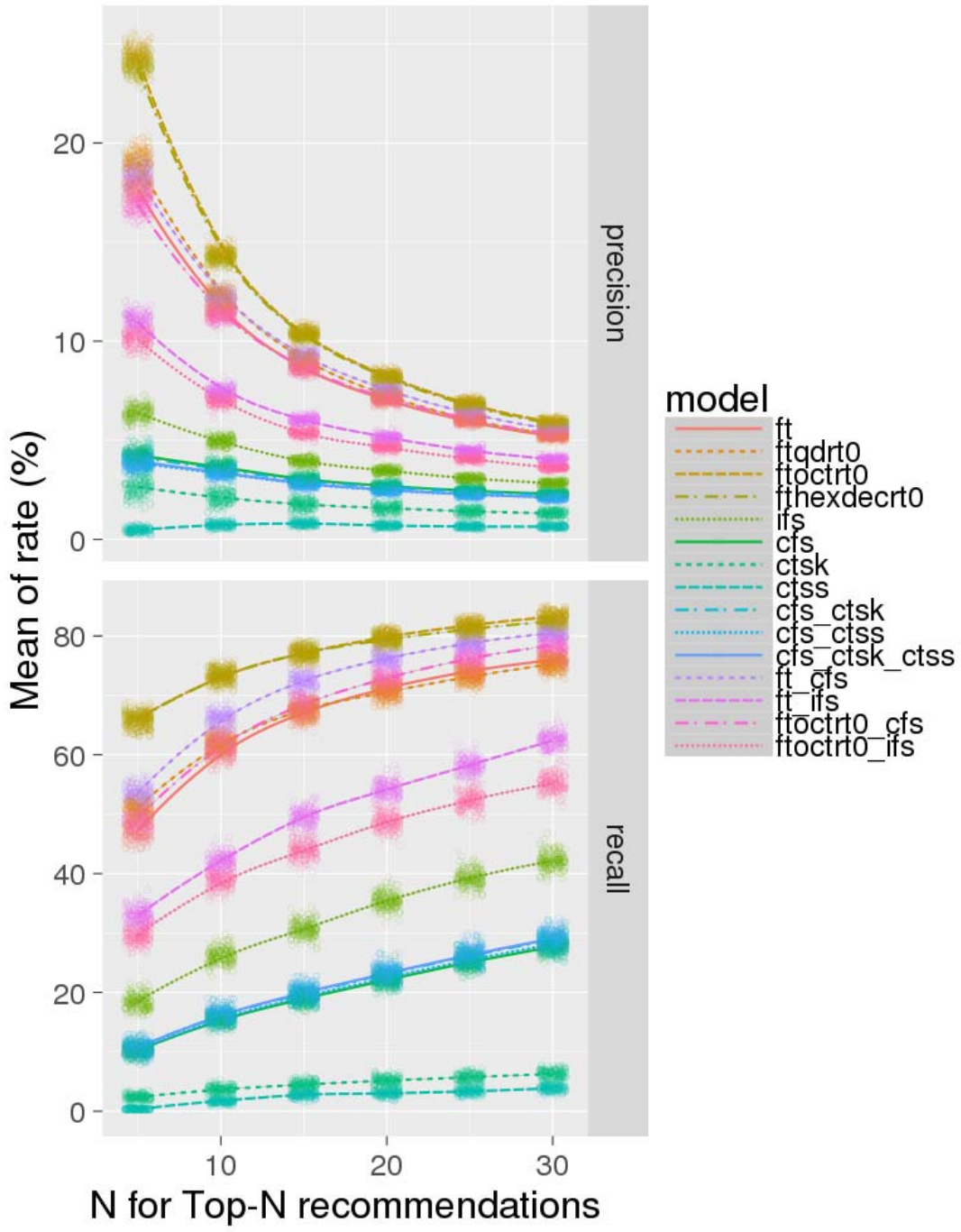


Figure 4.24: The precisions and recalls of the top-N recommendations for models built from data of year 2012 queried by data of year 2014

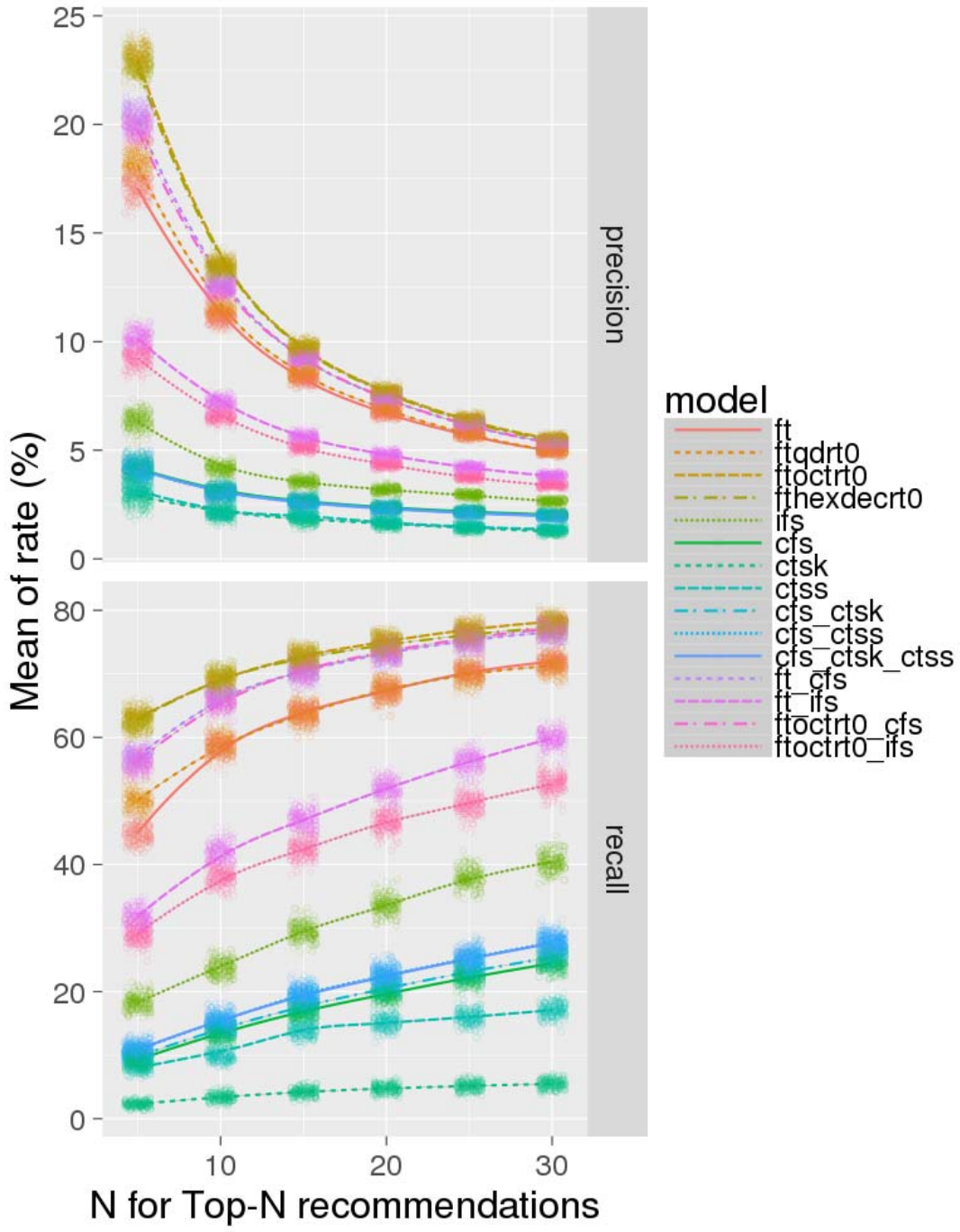


Figure 4.25: The precisions and recalls of the top-N recommendations for models built from data of year 2011 queried by data of year 2014

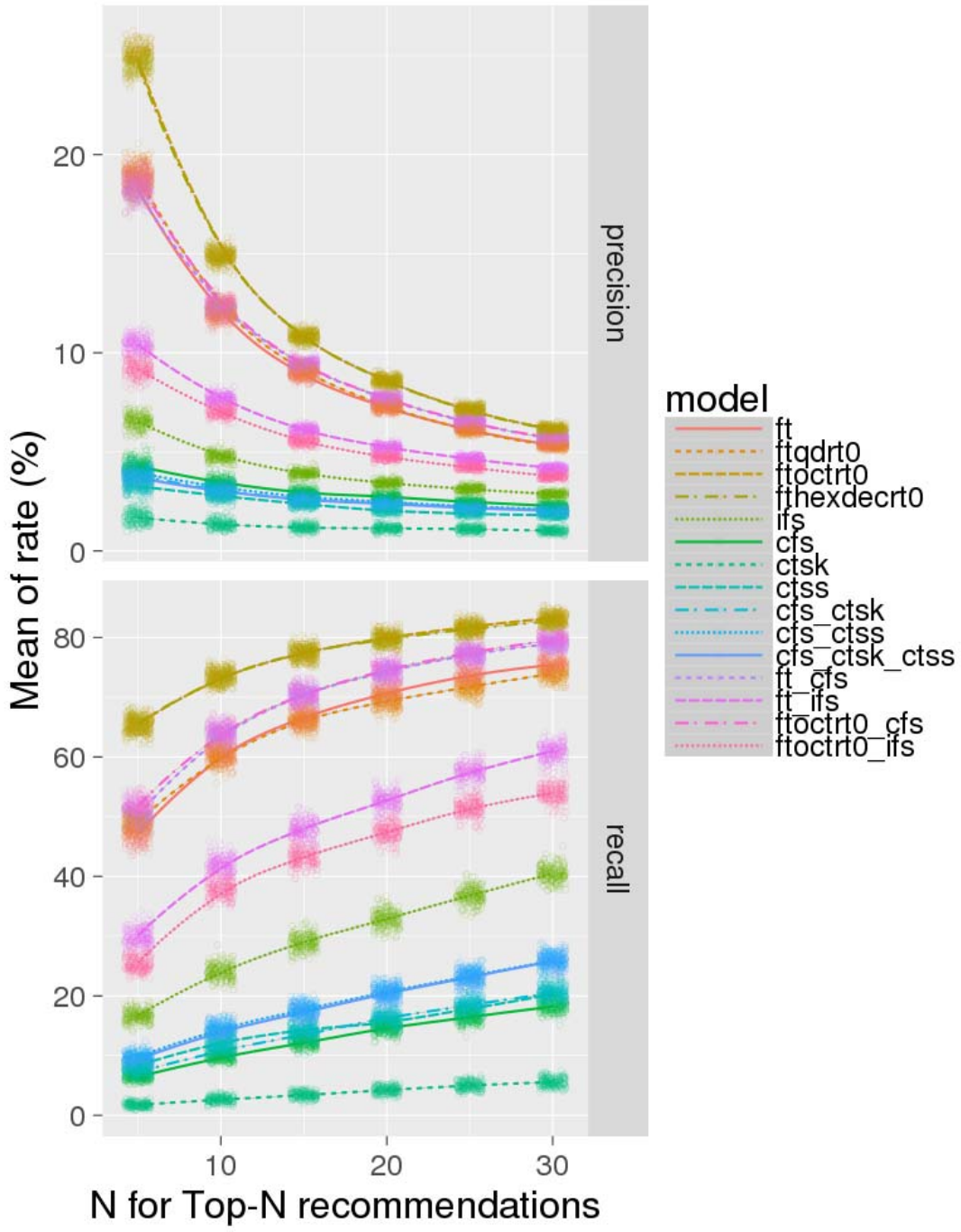


Figure 4.26: The precisions and recalls of the top-N recommendations for models built from data of year 2013 queried by data of year 2015

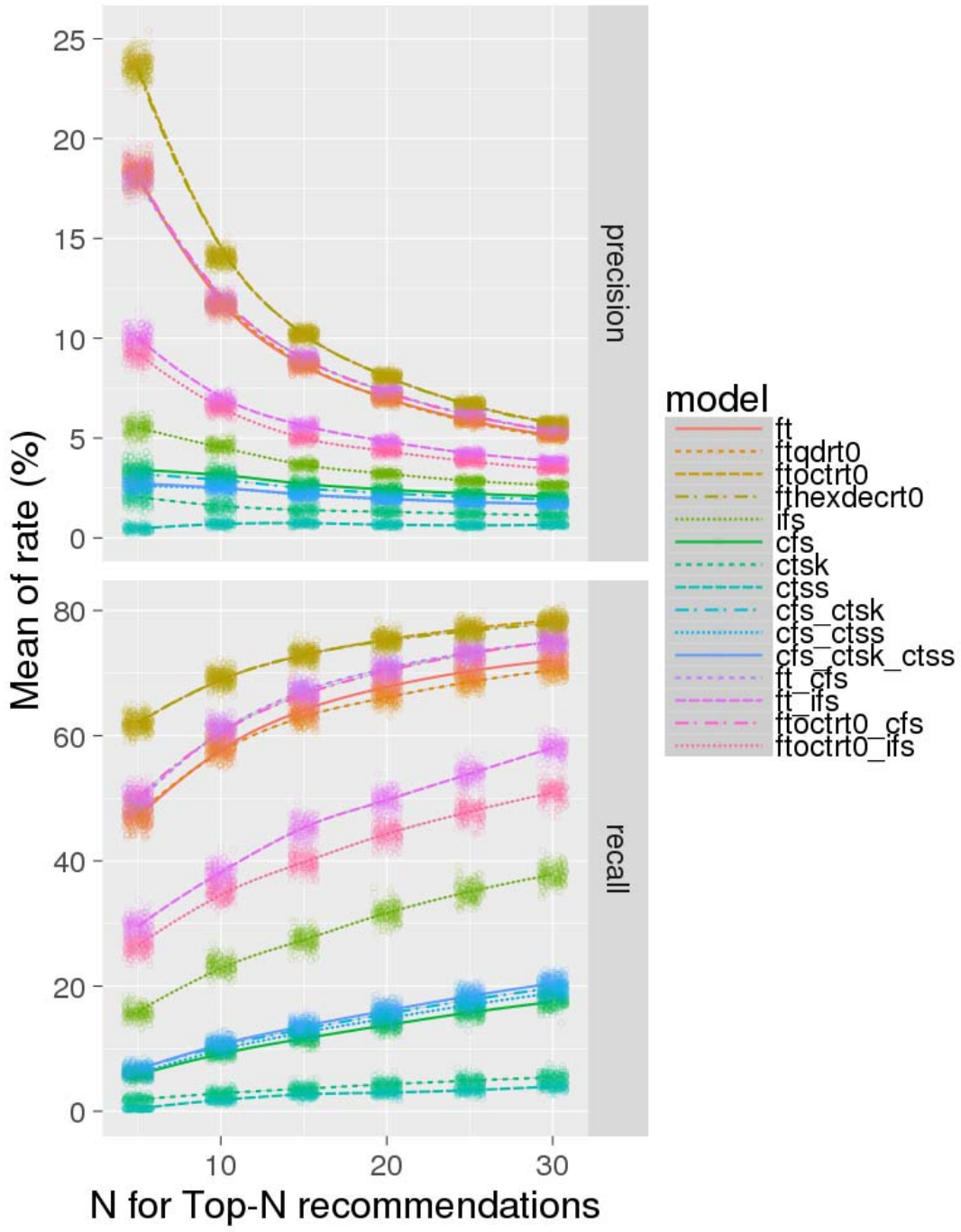


Figure 4.27: The precisions and recalls of the top-N recommendations for models built from data of year 2012 queried by data of year 2015

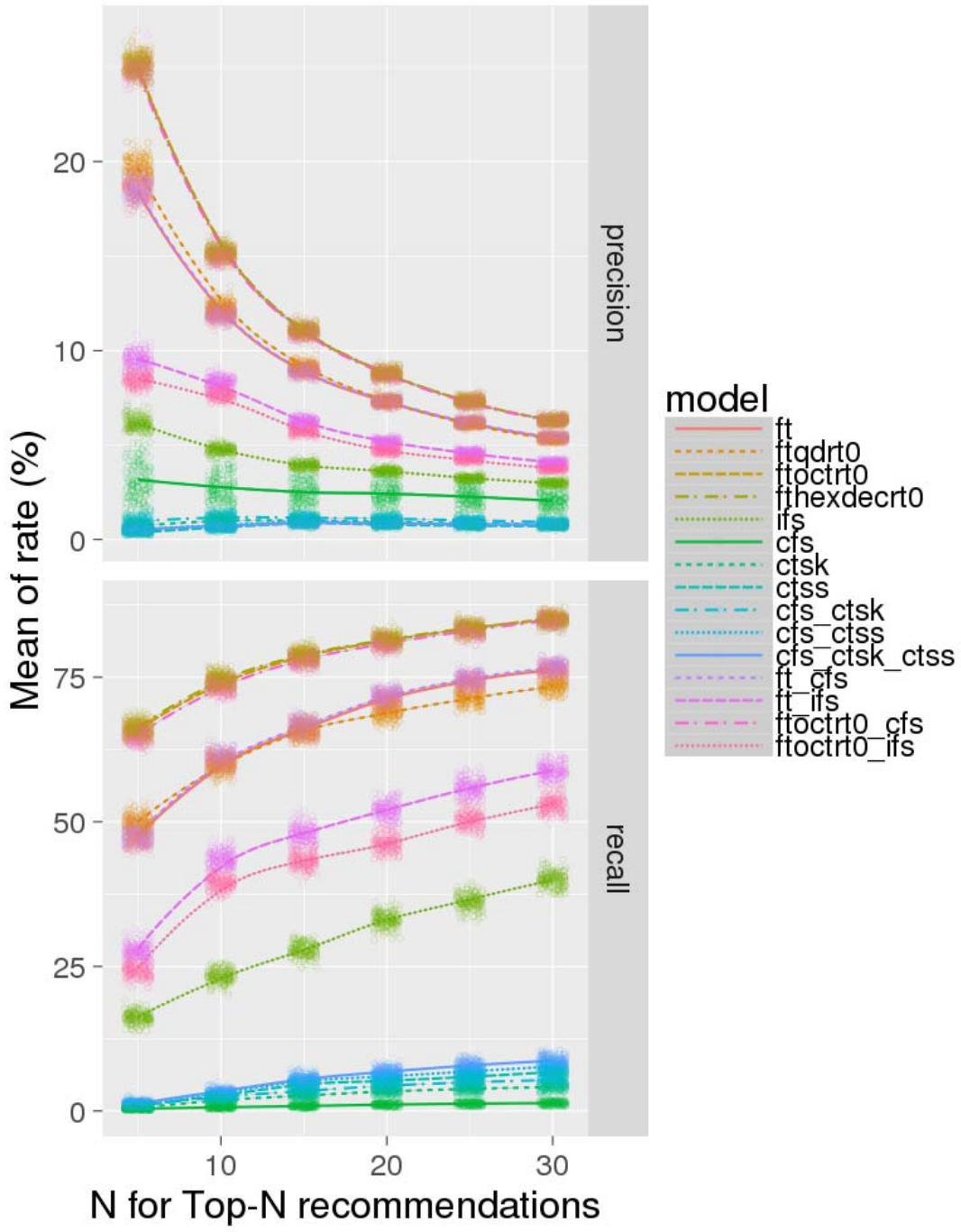


Figure 4.28: The precisions and recalls of the top-N recommendations for models built from data of year 2014 queried by data of year 2016

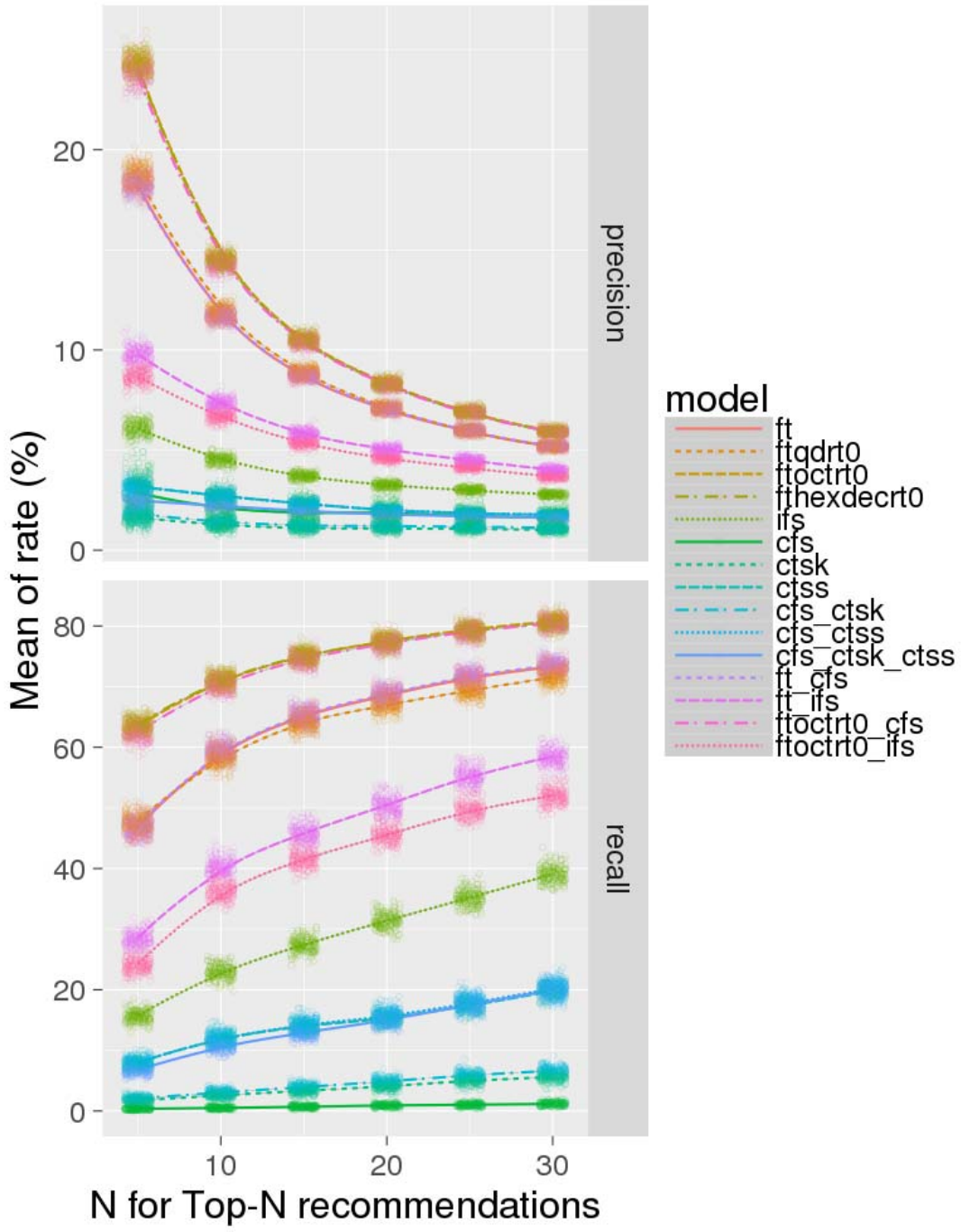


Figure 4.29: The precisions and recalls of the top-N recommendations for models built from data of year 2013 queried by data of year 2016

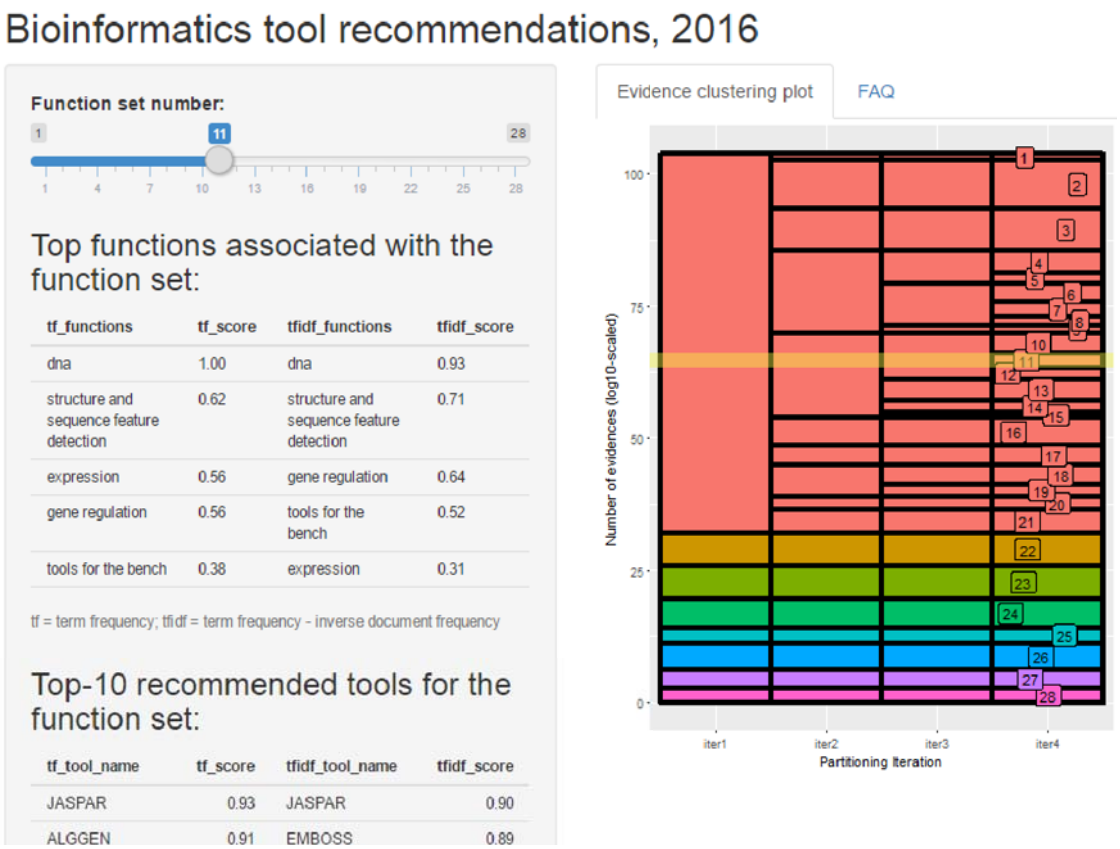


Figure 4.30: A screenshot of the tool recommender for function sets derived from data of 2016

CHAPTER V
CONCLUSION AND FUTURE WORK

This research gathered a list of known tools with associated PMIDs and identified a large corpus of usage evidences through crosscitations to the PMIDs. From the data, 90% of the tools were only publicized once. The result implies a high rate of tool discontinuity. The number of tools grew exponentially. The data suggested the current rate of practice change as two years. Keeping up with the changes requires continuous resource monitoring and adaptation. To faciliate the task, tool recommendation models were developed to best mimic the choices in the scholar community. The model fitting process inferred how the choices were made. Overall, tool functionalities remained the dominant influential factor. The tool popularity plays a significant role as well. Analysis patterns over the years, i.e. the converged function sets, were derived through the clustering method developed in this research. Changes in the field are reflected through the change in these analysis patterns. Tool recommendations made for each of the patterns can be compared between recommendation models learnt from differing years to monitor the change in tool preferences. A recommender was built to illustrate one of the many ways the knowledge can be used. Though built on a limited number of indexed tools, the recommendation list appears to be reflective of the community's status-quo.

Many challenges are still awaiting for future researches, e.g. the appropriate user query interface and the output format of the results. Informative output formats can lead to future works in automating the tool setup process. In this research, usage instances published in the scholar community is considered credible. Greater credibility standards such as including only publications published in journals with certain levels of impact factors can improve the quality of the recommendations. How the recommender can be made openly assessible and remain up-to-date in the long run also needs to be studied and be taken into action. Moreover, future improvements in tool indexing

methodologies are needed to maximize the usefulness of the work developed during this research.

REFERENCES

[1] National Institute of Health Maryland, USA *An illustration of the DNA double helix structure*.

[2] Stein, L. D. (2010). The case for cloud computing in genome informatics. *Genome Biol*, 11(5), 207.

[3] National Institute of Health Maryland, USA (2000). *NIH Working Definition of Bioinformatics and Computational Biology*.

[4] Xiong, J. (2006). *Essential bioinformatics*. Cambridge University Press, New York.

[5] Hogeweg, P. (2011). The roots of bioinformatics in theoretical biology. *PLoS Comput Biol*, 7(3), e1002021.

[6] Hagen, J. B. (2000). The origins of bioinformatics. *Nat Rev Genet*, 1(3), 231–236.

[7] Mount, D. (2004). *Bioinformatics : sequence and genome analysis*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

[8] Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48(3), 443–453.

[9] Nirenberg, M., Caskey, T., Marshall, R., Brimacombe, R., Kellogg, D., Doctor, B., Hatfield, D., Levin, J., Rottman, F., Pestka, S., Wilcox, M., and Anderson, F. (1966). The RNA code and protein synthesis. *Cold Spring Harbor Symposia on Quantitative Biology*, 31(0), 11–24.

[10] CRICK, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258), 561–563.

[11] Gibson, T. A. (2012). The roots of bioinformatics in ISMB. *PLoS Comput Biol*, 8(8), e1002679.

[12] Heather, J. M. and Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1), 1–8.

[13] Lewitter, F., Rebhan, M., Richter, B., and Sexton, D. (2009). The need for centralization of computational biology resources. *PLoS Comput Biol*, 5(6), e1000372.

[14] Cannata, N., Merelli, E., and Altman, R. B. (2005). Time to organize the bioinformatics resourceome. *PLoS Comp Biol*, 1(7), e76.

[15] Brazas, M. D., Yim, D., Yeung, W., and Ouellette, B. F. F. (2012). A decade of web server updates at the bioinformatics links directory: 2003-2012. *Nucleic Acids Research*, 40(W1), W3–W12.

[16] Chen, Y.-B., Chattopadhyay, A., Bergen, P., Gadd, C., and Tannery, N. (2007). The online bioinformatics resources collection at the university of pittsburgh health sciences library system–a one-stop gateway to online bioinformatics databases and software tools. *Nucleic Acids Research*, 35(Database), D780–D785.

[17] de la Calle, G., García-Remesal, M., Chiesa, S., de la Iglesia, D., and Maojo, V. (2009). BIRI: a new approach for automatically discovering and indexing available public bioinformatics resources from the literature. *BMC Bioinformatics*, 10(1), 320.

[18] de la Calle, G., García-Remesal, M., Nkumu-Mbomio, N., Kulikowski, C., and Maojo, V. (2012). e-MIR2: a public online inventory of medical informatics resources. *BMC Med Inform Decis Mak*, 12(1).

[19] Jon Ison, K. R. and Hervé Ménager, e. a., Matúš Kalaš (2015). Tools and data services registry: a community effort to document bioinformatics resources. *Nucleic Acids Res*, 44(D1), D38–D47.

[20] V. J. Henry, A.-S. P. B. J. G. A. D., A. E. Bandrowski (2014). OMICtools: an informative directory for multi-omic data analysis. *Database*, 2014(0), bau069–bau069.

[21] Eales, J. M., Pinney, J. W., Stevens, R. D., and Robertson, D. L. (2008). Methodology capture: discriminating between the ”best” and the rest of community practice. *BMC Bioinformatics*, 9(1), 359.

[22] Zhang, G., Ding, Y., and Milojević, S. (2013). Citation content analysis (CCA): A framework for syntactic and semantic analysis of citation content. *J Am Soc Inf Sci Tec*, 64(7), 1490–1503.

[23] Duck, G., Nenadic, G., Brass, A., Robertson, D. L., and Stevens, R. (2014). Extracting patterns of database and software usage from the bioinformatics literature. *Bioinformatics*, 30(17), i601–i608.

[24] Schaeffer, S. E. (2007). Graph clustering. *Computer Science Review*, 1(1), 27–64.

[25] Rosvall, M., Axelsson, D., and Bergstrom, C. T. (2009). The map equation. *The European Physical Journal Special Topics*, 178(1), 13–23.

[26] Li, S., Karatzoglou, A., and Gentile, C. (2016). Collaborative filtering bandits. *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval - SIGIR '16*.

[27] Bu, J., Shen, X., Xu, B., Chen, C., He, X., and Cai, D. (2016). Improving collaborative recommendation via user-item subgroups. *IEEE Trans. Knowl. Data Eng.*, 28(9), 2363–2375.

[28] Manning, C. D., Raghavan, P., and Schtze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

[29] Porter, M. F. (1997). Readings in information retrieval, chapter. An Algorithm for Suffix Stripping, pp. 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[30] R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

[31] Ooms, J., James, D., DebRoy, S., Wickham, H., and Horner, J. (2016). *RMySQL: Database Interface and 'MySQL' Driver for R*. R package version 0.10.9.

[32] Wickham, H. and Francois, R. (2016). *dplyr: A Grammar of Data Manipulation*. R package version 0.5.0.

[33] Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

[34] Huang, A. and Hoonlor, A. (2016). A multi-layer graph analytics to identify bioinformatics tool usage practices from tool directories and pubmed indexed

cross-citations. *The 20th International Computer Science and Engineering Conference 2016*, dec.

BIOGRAPHY

NAME	Miss. Angkana Huang
DATE OF BIRTH	15 June 1986
PLACE OF BIRTH	Bangkok, Thailand
INSTITUTIONS ATTENDED	Chulalongkorn University, 2004–2008
	Bachelor of Industrial Design (Ceramics)
	Mahidol University, 2015–2016
	Master of Science (Computer Science)
	Mahidol University
POSITION	Data Analyst
	Department of Virology, USAMD-AFRIMS
	315/6 Rajvithi Road, Bangkok, Thailand
E-MAIL	AngkanaH@afrims.org