

PGMHD: A Scalable Probabilistic Graphical Model for Massive Hierarchical Data Problems

Khalifeh AlJadda*, Mohammed Korayem†, Camilo Ortiz‡, Trey Grainger‡, John A. Miller* and William S. York§

*Department of Computer Science, University of Georgia, Athens, Georgia

Email: aljadda@uga.edu, jam@cs.uga.edu

† School of Informatics and Computing, Indiana University, Bloomington, IN

Email: mkorayem@cs.indiana.edu

‡CareerBuilder, Norcross, GA

Email: camilo.ortiz@careerbuilder.com, trey.grainger@careerbuilder.com

§Complex Carbohydrate Research Center, University of Georgia, Athens, Georgia

Email: will@ccrc.uga.edu

Abstract—In the big data era, scalability has become a crucial requirement for any useful computational model. Probabilistic graphical models are very useful for mining and discovering data insights, but they are not scalable enough to be suitable for big data problems. Bayesian Networks particularly demonstrate this limitation when their data is represented using few random variables with a massive set of outcome values for each of them. With hierarchical data - data that is arranged in a treelike structure with several levels - one would expect to see hundreds of thousands or millions of values distributed over even just a small number of levels. When modeling this kind of hierarchical data across large data sets, Bayesian networks become unsuitable for representing the probability distributions for the following reasons: i) each level represents a single random variable with hundreds of thousands of values, ii) the number of levels is usually small, so there are also few random variables, and iii) the structure of the network is predefined since the dependency is modeled top-down from each parent to each of its child nodes. In this paper we propose a scalable probabilistic graphical model to overcome these limitations for massive hierarchical data. We believe the proposed model will lead to an easily-scalable, more readable, and expressive implementation for problems that require probabilistic-based solutions for massive amounts of hierarchical data. We successfully applied this model to solve two different challenging probabilistic-based problems on massive hierarchical data sets for different domains, namely, bioinformatics and latent semantic discovery over search logs.

scalability issues that arise when they are applied to massive data sets.

Massive data sets often exhibit hierarchical properties, where data can be divided into several levels arranged in tree-like structures. Data items in each level depend on or are influenced by only the data items in the immediate upper level. For this kind of data the most appropriate PGM to represent the probability distribution would be a Bayesian network, since the dependencies in this kind of data are not bidirectional. A Bayesian network is appropriate when it can provide a concise representation of a large probability distribution where the joint probability cannot be efficiently handled using traditional techniques such as tables and equations [5]. Such a scenario is not the case with massive hierarchical data, as traditional Bayesian networks become infeasible to use. For example, in the glycan ontology "Glyco", [17] describes 1300 glycan structures whose theoretical tandem mass spectra (MS) can be predicted by GlycoWorkbench [3]. If the maximum of cleavages is set to two and the number of cross-ring cleavages is set to one, the theoretical MS^2 spectrum contains 2,979,334 ions, which themselves can be fragmented to form tens of millions of ions in MS^3 . To represent this data set of only two levels of the MS data using a Bayesian network it will require two nodes, MS^1 and MS^2 , with a single path $MS^1 \rightarrow MS^2$ while the conditional probability table (CPT) for the node MS^2 will contain 3,873,134,200 ($2,979,334 \times 1300$) entries.

I. INTRODUCTION

Probabilistic graphical models (PGM) refer to a family of techniques that merge concepts from graph structures and probability models [16]. They represent the conditional dependencies among sets of random variables [9]. In the age of big data, PGMs can be very useful for mining and extracting insights from large-scale and noisy data. The major challenges that PGMs face in this emerging field are the scalability and the restriction that they can only be applied on domains of limited size (e.g. propositional domain) [8], [4]. Some extensions have already been proposed to address these challenges, such as hierarchical probabilistic graphical models (HPGM) which aim to extend the PGM to work with more structured domains [8], [6]. The focus of these models is to make Bayesian networks applicable to structured domains, but they do not solve the

The main contributions of this paper are as follows. First, we propose an efficient and scalable probabilistic-based model for massive hierarchical data (PGMHD). Second, we successfully apply PGMHD to the bioinformatics domain described above in which we automatically classify and annotate high-throughput mass spectrometry data. Finally, we successfully apply this model to large-scale latent semantic discovery by using 1.6 billion search log entries provided by CareerBuilder.com within a Hadoop Map/Reduce framework.

II. BACKGROUND

Graphical models can be classified into two major categories: (1) directed graphical models (the focus of this paper), which are often referred to as Bayesian networks, or belief networks, and (2) undirected graphical models which are often

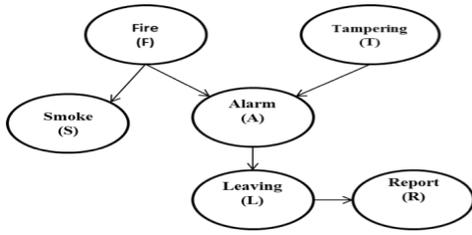


Fig. 1. Bayesian Network [5]

referred to as Markov Random Fields, Markov networks, Boltzmann machines, or log-linear models [11]. Probabilistic graphical models (PGMs) consist of both graph structure and parameters. The graph structure represents a set of conditionally independent relations for the probability model, while the parameters consist of the joint probability distributions [16].

A Bayesian network is a concise representation of a large probability distribution to be handled using traditional techniques such as tables and equations [5]. The graph of a Bayesian network is a directed acyclic graph (DAG) [9]. A Bayesian network consists of two components: a DAG representing the structure (as shown in Figure 1), and a set of conditional probability tables (CPTs). Each node in a Bayesian network must have a CPT which quantifies the relationship between the variable represented by that node and its parents in the network. Completeness and consistency are guaranteed in a Bayesian network since there is only one probability distribution that satisfies the Bayesian network constraints [5]. The constraints that guarantee a unique probability distribution are the numerical constraints represented by CPT and the independence constraints represented by the structure itself. The independence constraints are shown in Figure 1. Each variable in the structure is independent of any other variables other than its parents, once its parents are known. For example, once the information about A is known, the probability of L will not be affected by any new information about F or T, so we call L independent of F and T once A is known.

Bayesian networks are widely used for modeling causality in a formal way, for decision-making under uncertainty, and for many other applications [5].

III. RELATED WORK

This section describes the most related work to the proposed model from different perspectives. First, we describe the related hierarchical probabilistic models, then we describe the current techniques used to automate the annotation of Mass Spectrometry (MS) data for glycomics, which is one of the scenarios that we use to test the proposed model. We close this section by describing how we applied the proposed model to discover the latent semantic similarity between keywords extracted from search logs for the purposes of building a semantic search system.

A. Probabilistic Graphical Models for Hierarchical Data

Probabilistic graphical models require simple domains [8]. To overcome this common limitation some extensions were proposed to extend those models to structured domains. In [8], the authors introduced a hierarchical Bayesian network which

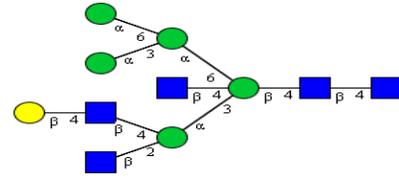


Fig. 2. Glycan structure in CFG format. The circles and squares represent the monosaccharides which are the building blocks of a glycan while the lines are the linkages between them

extends the expressiveness of a regular Bayesian network by allowing a node to represent an aggregation of simpler types which enables the modeling of complex hierarchical domains. The main idea is to use a small number of hidden variables as a compressed representation for a set of observed variables with the following restrictions: 1) Any parent of a variable should be in the same or immediate upper layer, and 2) At most one parent from the immediate upper layer is allowed for each variable.

So, the idea is mainly to compress the observed data. Although hierarchical Bayesian network models extended the regular Bayesian network to represent structured domains, they have not been able to solve the issue of the scalability of Bayesian networks for massive amounts of hierarchical data.

B. Automated Annotation of Mass Spectrometry Data for Glycomics

One use case of the proposed model is the automated annotation of Mass Spectrometry (MS) data for glycomics. Glycans (Figure 2) are the third major class of biological macro-molecules besides nucleic acids and proteins [1]. Glycomics refers to the scientific attempts to characterize and study glycans, as defined in [1] or an integrated systems approach to study structure-function relationships of glycans as defined in [14]. The importance of this emerging field of study is clear from the accumulated evidence for the roles of glycans in cell growth and metastasis, cell-cell communication, and microbial pathogenesis. Glycans are more diverse in terms of chemical structure and information density than nucleic acids and proteins [14]. Unlike proteins, which can be represented as linear sequences, glycans are branched structures that are often represented as trees. Thus glycan identification, which is a proven NP-hard problem [15], is much more difficult than protein identification.

Although MS has become the major analytical technique for glycans, no general method has been developed for the automated identification of glycan structures using MS and tandem MS data. The relative ease of peptide identification using tandem MS is mainly due to the linear structure of peptides and the availability of reliable peptide sequence databases. In proteomic analyses, a mostly complete series of high abundance fragment ions is often observed. In such tandem mass spectra, the mass of each amino acid in the sequence corresponds to the mass difference between two high-abundance peaks, allowing the amino acid sequence to be deduced. In glycomics MS data, ion series are disrupted by the branched nature of the molecule, significantly complicating the extraction of sequence

information. In addition, groups of isomeric monosaccharides commonly share the same mass, making it impossible to distinguish them by MS alone. Databases for glycans exist but are limited, minimally curated, and suffer badly from pollution from glycan structures that are not produced in nature or are irrelevant to the organism of study. Several algorithms have been developed in attempts to semi-automate the process of glycan identification by interpreting tandem MS spectra, including CartoonistTwo [7], GlycoWork-bench [3], and SimGlycan [2] (commercially available from Premier Biosoft). However, each of these programs produces incorrect results when using polluted databases to annotate large MSⁿ datasets containing hundreds or thousands of spectra. Inspection of the current literature indicates that machine learning and data mining techniques have not been used to resolve this issue, although they have a great potential to be successful in doing so. PGMHD attempts to employ machine learning techniques (mainly probabilistic-based classification) to find a solution for the automated identification of glycans using MS data.

C. Semantic Similarity

Semantic similarity, which is a metric that is defined over documents or terms in which the distance between them reflects the likeness of their meaning [10], is well defined in Natural Language Processing (NLP) and Information Retrieval (IR) [13]. The major techniques to calculate semantic similarity are Pointwise Mutual Information (PMI) and Latent Semantic Analysis (LSA), though PMI outperform LSA on mining the web for synonyms [18]. We applied the proposed PGMHD model to discover related search terms by measuring probabilistic-based semantic similarity between those search terms.

IV. MODEL STRUCTURE

In this section, we discuss the proposed PGMHD model, describing the computation of a probabilistic-based classification, calculation of a probabilistic-based similarity score, and the progressive learning aspects of the model.

Consider multi-level directed graph $G = (V, A)$ where V and $A \subset V \times V$ denote the sets of nodes and arcs, respectively, such that:

- 1) V is partitioned into m levels L_0, \dots, L_{m-1} such that $V = \cup_{i=0}^{m-1} L_i$, and $L_i \cup L_j = \emptyset$ for $i \neq j$.
- 2) The arcs in A only connect one level to the next, i.e., if $a \in A$ then $a \in L_{i-1} \times L_i$ for some $i = 1, \dots, m-1$.
- 3) An arc $a = (v_{i-1}, v_i) \in L_{i-1} \times L_i$ represents the dependency of v_i with its parent v_{i-1} , $i = 1, \dots, m-1$. Moreover, let $\text{pa} : V \rightarrow \mathcal{P}(V)$ be the function that maps every node to its parents, i.e.,

$$\text{pa}(v) = \{w : (w, v) \in A\} \quad \forall v \in V.$$

- 4) The nodes in each level L_i represent all the possible outcomes of a finite discrete random variable, namely X_i , $i = 1, \dots, m-1$.

Note that the nodes in the first level L_0 can be seen as root nodes and the ones in L_{m-1} as leaves. Also, an observation x

in our probabilistic model is an outcome of a random variable, namely $X \in L_0 \times \dots \times L_{m-1}$, defined as

$$X = (X_0, X_1, \dots, X_m), \quad (1)$$

which represents a path from L_0 to L_{m-1} such that $(X_{i-1}, X_i) \in A$.

In addition, we assume that there are n observations of X , namely x^1, \dots, x^n , and let $f : V \times V \rightarrow \mathbb{N}$ be a frequency function defined as

$$f(a) = \left| \left\{ x^j : (x_{i-1}^j, x_i^j) = a, i \in \{1, \dots, m-1\}, j \in \{1, \dots, n\} \right\} \right|,$$

for every $a \in V \times V$. Moreover, these latter n observations are the ones used to train our model, so that $f(a) > 0$ for every $a \in A$.

It should be observed that the proposed model can be seen as a special case of a Bayesian network by considering a network consisting of a single directed path with m nodes. However, we believe that a leveled directed graph that explicitly defines one node per outcome of the random variables (as described above): i) leads to an easily scalable (and distributable) implementation of the problems we consider; ii) improves the readability and expressiveness of the implemented network; and iii) simplifies and facilitates the training of the model.

A. Probabilistic-based Classification

Given an outcome at level $i \in \{1, \dots, m-1\}$, namely $v \in L_i$, we calculate the *classification score* $\text{Cl}_i(w|v)$ of v to the parent outcome $w \in L_{i-1}$ by estimating the conditional probability $P(X_{i-1} = w | X_i = v)$ as follows

$$\begin{aligned} P(X_{i-1} = w | X_i = v) &= \frac{P(X_i = v | X_{i-1} = w) \cdot P(X_{i-1} = w)}{P(X_i = v)} \\ &\approx \frac{\left(\frac{f(w, v)}{\text{Out}(w)} \right) \cdot \left(\frac{\text{Out}(w)}{n} \right)}{\left(\frac{\text{In}(v)}{n} \right)} = \frac{f(w, v)}{\text{In}(v)} =: \text{Cl}_i(w|v). \end{aligned}$$

where

$$\text{In}(v) := \sum_{u \in \text{pa}(v)} f(u, v), \quad \forall v \in V,$$

and

$$\text{Out}(w) := \sum_{u : (w, u) \in A} f(w, u), \quad \forall w \in V.$$

B. Probabilistic-based Similarity scoring

Fix a level $i \in \{1, \dots, m-1\}$, and let $X, Y \in L_0 \times \dots \times L_{m-1}$ be identically distributed random variables as in (1). We define the *probabilistic-based similarity score* between two outcomes $x_i, y_i \in L_i$ by computing the conditional joint probability

$$\text{CO}_i(x_i, y_i) := P(X_i = x_i, Y_i = y_i | X_{i-1} \in \text{pa}(x_i), Y_{i-1} \in \text{pa}(y_i))$$

as

$$\text{CO}_i(x_i, y_i) = \prod_{v \in \text{pa}(x_i)} p_i(v, x_i) \cdot \prod_{v \in \text{pa}(y_i)} p_i(v, y_i),$$

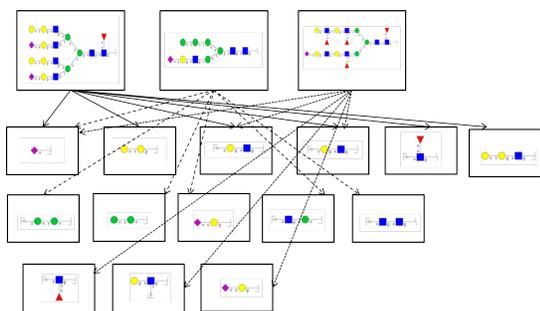


Fig. 3. PGMHD for tandem MS data. The root nodes are the glycans that annotate the peaks at MS¹ level, while the level 2 nodes are the glycan fragments that annotate the peaks at MS² level and the edges represent dependency associating the glycans with their MS² fragments.

where $p_i(v, w) = P(X_{i-1} = v, X_i = w)$ for every $(v, w) \in L_{i-1} \times L_i$. Recalling that n is the number of observations, we can naturally estimate the probabilities $p_i(v, w)$ with $\hat{p}(v, w)$ defined as

$$\hat{p}(v, w) := \frac{f(v, w)}{n}.$$

Hence, we can obtain the related outcomes of $x_i \in L_i$ (at level i) by finding all the $w \in L_i$ with a large estimated probabilistic-based similarity score $\text{CO}_i(x_i, w)$.

C. Progressive Learning

PGMHD is designed to allow progressive learning¹. Progressive learning is a learning technique that allows a model to learn gradually over time. Training data does not need to be given at one time to the model. Instead, the model can learn from any available data and integrate the new knowledge with the represented one. This learning technique is very attractive in the big data age for the following reasons:

- 1) Training the model does not require processing all data upfront
- 2) It can easily learn from new data without the need to re-include the previous training data in the learning.
- 3) The training session can be distributed instead of doing it in one long-running session.
- 4) Recursive learning allows the results of the model to be used as new training data, provided they are judged to be accurate by the user.

V. EXPERIMENTAL RESULTS

PGMHD can be used for different purposes once it is built and trained. PGMHD can be used to predict the class from level l for the observations of random variables at level $l+1$. For example, in the annotation of the MS data, PGMHD is used to predict the best Glycan at level MS¹ to annotate a spectrum by evaluating the annotated peaks at level MS² with probability scores that represent how well the selected glycan correlates to the manually curated annotations that were used to train the model.

Scan #	Peak Charge	Peak Intensity	Peak m/z	Cartoon	Feature m/z	Glycan Id
117	3	144887.0	1085.923		1085.5136	GOG166
118	-1	249107.0	812.4		812.0444	GOG120
119	2	79236.5	1023.9		1023.7372	GOG516

Fig. 4. MS1 annotation using GELATO. Scan is the ID number of the scan in the MS file, peak charge is the charge state of that peak in the MS file, peak intensity represents the abundance of an ion at that peak, peak m/z is the mass over charge of the given peak, cartoon is the annotation of that peak (glycan) in CFG format, feature m/z is the mass over charge for the glycan, and glycanID is the ID of the glycan in the Glycan Ontology(Glyco).

Peak Intensity	Peak m/z	Cartoon	Feature m/z	Glycan Id
13.1097	398.3249		398.1693	GOG166
5.2044	445.3975		445.1952	GOG166
13.2528	472.3444		472.2061	GOG166
13.2528	472.3444		472.2061	GOG166
13.2528	472.3444		472.2061	GOG166
8.7515	474.3912		474.2217	GOG166
2.8145	690.4586		690.3215	GOG166
2.8145	690.4586		690.3215	GOG166

Fig. 5. Fragments of Glycan GOG166 at the MS² level. Each ion observed in MS¹ is selected and fragmented in MS² to generate smaller ions, which can be used to identify the glycan structure that most appropriately annotates the MS¹ ion. Theoretical fragments of the glycan structure that had been used to annotate the MS¹ spectrum are used to annotate the corresponding MS² spectrum.

A. PGMHD to automate the MS annotation

This model is well suited for representing MS data. We recently implemented the Glycan Elucidation and Annotation Tool (GELATO), which is a semi-automated MS annotation tool for glycomics integrated within our MS data processing framework called GRITS (<http://www.grits-toolbox.org/>). Figures 4, and 5 show screen shots from GELATO for annotated spectra. Figure 4 shows the MS profile level and Figure 5 shows the annotation of MS² peaks using fragments of a selected candidate glycan for annotation of the MS¹ data.

To represent the data shown in these figures using the proposed model, a top-layer node is assigned to each row in the MS profile table, which corresponds to the MS¹ data. Then, for each row in the MS² tables, a unique node is created and connected with its parent node using a directed edge from the parent node (at the MS profile layer) to the child node (at the MS² layer). Each top-layer node stores a value representing how frequently a parent has been seen in the training data. However, each child node in the MS² layer has more than one parent. The edge's weight represents the co-occurrence frequency between a child and a parent, storing frequencies rather than probabilities facilitates progressive learning. The child node stores the total frequency of observing that child regardless of the identity of its parents. The combined frequency data makes it possible to design a progressive learning algorithm that can extract information from massive data sets. Figure 3 shows the PGMHD for the given MS data in these figures. As shown in the model, two layers are created: one

¹The progressive learning approach for PGMHD can be found on <http://www.aljadda.net/ProgLearn.pdf>.

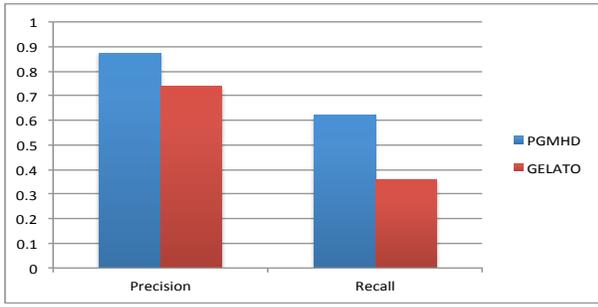


Fig. 6. Average precision and recall of PGMHD and GELATO

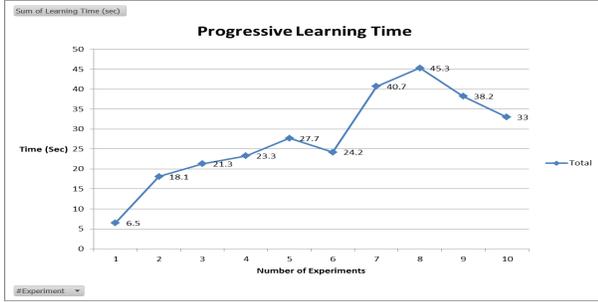


Fig. 7. Progressive Learning Time Over Different Experiments

for the MS¹ level and a second one for the MS² level. Several different nodes at the MS¹ level can be annotated with the same fragment ion at the MS² level, so MS² nodes can have several parents. The frequency values are not shown because of space constraints.

Experiments were performed using MS data collected from stem cell samples. The size of this data set is 1,746,278 peaks distributed over 1713 MS scans from 10 MS experiments. Figure 7 shows the learning time using the progressive learning technique. In this test we introduced one new experiment at a time to the model for training, and we recorded the total time required to train the model. These performance results demonstrate how efficiently the progressive learning works with PGMHD.

To test the accuracy of PGMHD, we trained the model by randomly selecting one of 10 available experiments, while the other 9 experiments were used to test the trained model by annotating the experiments' peaks using PGMHD. The baseline in our evaluation was the annotations generated by the commercial tool SimGlycan. Figure 6 shows the average precision and recall for PGMHD compared to the average precision and recall of GELATO using the same dataset of 1,746,278 peaks distributed over 10 MS experiments.

B. PGMHD for latent semantic discovery over Hadoop

We also implemented a version of PGMHD over Hadoop [12] to be used for latent semantic discovery between users' search terms extracted from search logs provided by CareerBuilder.com.

1) *Problem Description*: CareerBuilder operates the largest job board in the U.S. and has an extensive and growing global presence, with millions of job postings, more than 60 million

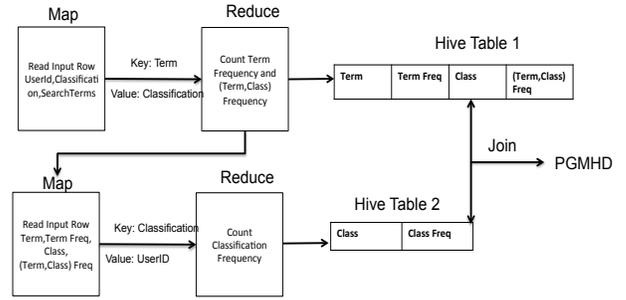


Fig. 8. PGMHD Over Hadoop

TABLE I. INPUT DATA TO PGMHD OVER HADOOP

UserID	Classification	Search Terms
user1	Java Developer	Java, Java Developer, C, Software Engineer
user2	Nurse	RN, Registered Nurse, Health Care
user3	.NET Developer	C, ASP, VB, Software Engineer, SE
user4	Java Developer	Java, JEE, Struts, Software Engineer, SE
user5	Health Care	Health Care Rep, HealthCare

actively-searchable resumes, over one billion searchable documents, and more than a million searches per hour. The search relevancy and recommendations team wanted to discover latent semantic relationships among the search terms entered by their users in order to build a semantic search engine that understands a user's query intent in order to provide more relevant results than a traditional keyword search engine. To tackle this problem, CareerBuilder cannot use typical synonym dictionaries since most of the keywords used in the employment search domain represent job titles, skills, and companies that would not be found in a traditional English dictionary. Additionally, CareerBuilder's search engine supports over a dozen languages, so they were in search of a model that is language-independent.

2) *PGMHD over Hadoop*: Given the search logs for all the users and the users' classifications as shown in Table I, PGMHD can represent this kind of data by placing the classes of the users as root nodes and placing the search terms for all the users in the second level as children nodes. Then, an edge will be formed linking each search term back to the class of the user who searched for it. The frequency of each search term (how many users search for it) will be stored in the node of that term, while the frequency of a specific search term searched for by users of a specific class (how many users belonging to that class searched for the given term) will be stored in the edge between the class and the term. The frequency of the root node is the summation of the frequencies of the edges that connect that root node with its children (Figure 9).

Figure 8 shows how PGMHD was implemented over Hadoop using Map/Reduce jobs and Hive tables. After we created PGMHD on Hadoop we calculated the probabilistic-based semantic similarity score between each pair of two terms with shared parents. The size of the data set we analyzed in this experiment is 1.6 billion search records. To decrease the noise in the given data set we applied a pre-filtering technique by removing any search term used by less than 10 distinct users. The final graph representing this data contains 1931 root nodes, 16,414 child nodes, and 439,435 edges.

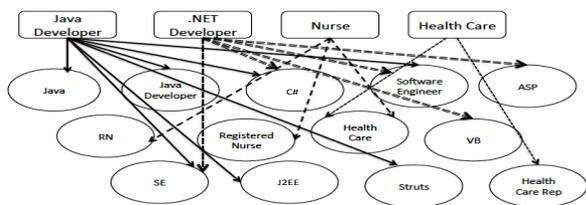


Fig. 9. PGMHD representing the search log data

TABLE II. PGMHD RESULTS FOR LATENT SEMANTIC DISCOVERY

Term	Related Terms
hadoop	big data, hadoop developer, obiee, java, python
registered nurse	rn registered nurse, rn, registered nurse manager, nurse, nursing, director of nursing
data mining	machine learning, data scientist, analytics, business intelligence, statistical analyst
solr	lucene, hadoop, java

3) Results of latent semantic discovery using PGMHD:

The experiment performing latent semantic discovery among search terms using PGMHD was run on a Hadoop cluster with 63 data nodes, each having a 2.6 GHZ AMD Opteron Processor with 12 to 32 cores and 32 to 128 GB RAM. Table II shows sample results of 10 terms with their top 5 related terms discovered by PGMHD. To evaluate the model's accuracy, we sent the results to data analysts at CareerBuilder who reviewed 1000 random pairs of discovered related search terms and returned the list with their feedback about whether each pair of discovered related terms was "related" or "unrelated". We then calculated the accuracy (precision) of the model based upon the ratio of the number of related results to the total number of results. The results show the accuracy of the discovered semantic relationships among search terms using the PGMHD model to be 0.80.

VI. CONCLUSION

Probabilistic graphical models are very important in many modern applications such as data mining and data analytics. A major issue with existing probabilistic graphical models is their scalability to handle large data sets, making this a very important area for research given the tremendous focus on big data due to the growing number of data points produced by modern computers systems and sensors. PGMHD is a probabilistic graphical model that attempts to solve the scalability problems with existing models for scenarios involving massive hierarchical data. PGMHD is designed to fit hierarchical data sets of any size, regardless of the domain to which the data belongs. In this paper, we present two experiments from different domains: one being the automated annotation of high-throughput mass spectrometry data in bioinformatics, and the other being latent semantic discovery using search logs from the largest job board in the U.S. The two use cases in which we tested PGMHD show that this model is robust and can scale from a few thousand entries to billions of entries, and can also run on a single computer (for smaller data sets), as well as in a parallelized fashion on a large cluster of servers (63 were used in our experiment).

ACKNOWLEDGMENT

The authors would like to deeply thank David Crandall from Indiana University, Kiyoko Aoki Kinoshita from Soka University, and Khaled Rasheed from University of Georgia for the valuable discussions and suggestions to improve this model. We would also like to thank Melody Porterfield and Rene Ranzinger from the Complex Carbohydrate Research Center (CCRC) at the University of Georgia for providing their help in the glycan annotation experiments. This work was supported by the National Institute of General Medical Sciences, a part of the National Institutes of Health, funding the National Center for Glycomics and Glycoproteomics (8P41GM103490).

REFERENCES

- [1] K. F. Aoki-Kinoshita. An introduction to bioinformatics for glycomics research. *PLoS computational biology*, 4(5):e1000075, 2008.
- [2] A. Apte and N. S. Meitei. Bioinformatics in glycomics: Glycan characterization with mass spectrometric data using singlycanâĎ. In *Functional Glycomics*, pages 269–281. Springer, 2010.
- [3] A. Ceroni, K. Maass, H. Geyer, R. Geyer, A. Dell, and S. M. Haslam. Glycworkbench: a tool for the computer-assisted annotation of mass spectra of glycans. *Journal of proteome research*, 7(4):1650–1659, 2008.
- [4] D. M. Chickering, D. Heckerman, and C. Meek. Large-sample learning of bayesian networks is np-hard. *The Journal of Machine Learning Research*, 5:1287–1330, 2004.
- [5] A. Darwiche. Bayesian networks. *Communications of the ACM*, 53(12):80–90, 2010.
- [6] S. Fine, Y. Singer, and N. Tishby. The hierarchical hidden markov model: Analysis and applications. *Machine learning*, 32(1):41–62, 1998.
- [7] D. Goldberg, M. Bern, B. Li, and C. B. Lebrilla. Automatic determination of o-glycan structure from fragmentation spectra. *Journal of proteome research*, 5(6):1429–1434, 2006.
- [8] E. Gyftodimos and P. A. Flach. Hierarchical bayesian networks: an approach to classification and learning for structured data. In *Methods and Applications of Artificial Intelligence*, pages 291–300. Springer, 2004.
- [9] T. Hamelryck. An overview of bayesian inference and graphical models. In *Bayesian Methods in Structural Bioinformatics*, pages 3–48. Springer, 2012.
- [10] S. Harispe, S. Ranwez, S. Janaqi, and J. Montmain. Semantic measures for the comparison of units of language, concepts or entities from text and knowledge base analysis. *arXiv preprint arXiv:1310.1285*, 2013.
- [11] M. I. Jordan et al. Graphical models. *Statistical Science*, 19(1):140–155, 2004.
- [12] C. Lam. *Hadoop in action*. Manning Publications Co., 2010.
- [13] R. Mihalcea, C. Corley, and C. Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, volume 6, pages 775–780, 2006.
- [14] R. Raman, S. Raguram, G. Venkataraman, J. C. Paulson, and R. Sasisekharan. Glycomics: an integrated systems approach to structure-function relationships of glycans. *Nature Methods*, 2(11):817–824, 2005.
- [15] B. Shan, B. Ma, K. Zhang, and G. Lajoie. Complexities and algorithms for glycan sequencing using tandem mass spectrometry. *Journal of bioinformatics and computational biology*, 6(01):77–91, 2008.
- [16] P. Smyth. Belief networks, hidden markov models, and markov random fields: A unifying view. *Pattern recognition letters*, 18(11):1261–1268, 1997.
- [17] C. J. Thomas, A. P. Sheth, and W. S. York. Modular ontology design using canonical building blocks in the biochemistry domain. *Frontiers in Artificial Intelligence and Applications*, 150:115, 2006.
- [18] P. D. Turney. Mining the web for synonyms: PMI-IR versus lsa on toefl. In *Proceedings of the 12th European Conference on Machine Learning*, EMCL '01, pages 491–502, London, UK, UK, 2001. Springer-Verlag.