Enhancing the performance of clustering algorithms using MapReduce computations*

*

Nhlahla Gershom Maluleke
Deprtment of Computer Science, School of IT
University Of Pretoria
Pretoria, South Africa
u13229908@tuks.co.za

***Abstract*—This document is a model and instructions for LaTeX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. *CRITICAL: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

This document is a model and instructions for LaTeX. Please observe the conference page limits.

II. Ease of Use

Big Data

Over the last two decades, data in various fields has increased in a very large scale. International Data Corporation(IDC) in 2011, has reported an overall volume of 1.8 ZB(Zettabyte) of both copied and created data in the world. This means that the overall data created and copied increased by nearly 5 times over the years. The recent advances within fields such as Information Technology(IT) has made the generation of more data easier. For example, social networking sites such as Twitter and Facebook have data that span over several terabytes. YouTube has been reported to have a similar size and it produces hundreds of gigabytes per minute.

Every minute, 72 hours on average of videos are uploaded by YouTube users. Facebook generates log data amounting to a sum of 10 petabytes a month. As a result, we are faced with a challenge of collecting and integrating enormous amounts of data from widely distributed sources of data. Other advances that further promote a sharp growth of data include the rapid growth of Internet of Things and cloud computing. Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on hand database management tools or traditional data processing applications. Under the explosive increase of data globally, the phrase Big data usually refers to enormous datasets. In comparison with normal or traditional datasets, this dataset is usually unstructured which makes it very difficult and challenging to analyse and interpret the datasets. Therefore, we need to establish techniques to analyse the data and get results within an acceptable amount of time.

Some of the benefits of Big data include the following, firstly it brings about new opportunities for data miners to discover new patterns and values. Secondly, after discovering these new values, Big data helps data miners to gain an in-depth understanding of hidden values within the data. Lastly, it incurs new challenges such as how to manage and organize large datasets. Big data decodes data that has been untouched in order to derive new insight which will be integrated with business operations. However, as data continues to grow exponentially, most of the current data mining techniques become obsolete.

Characteristics of Big data

Big data is described by three characteristics which are known as the 3Vs of Big data. These characteristics are volume, velocity and variety. Volume refers the large amounts of data that is produced endlessly. Initially, the actual storage of this data was very challenging due to storage costs being high. Over the years, storage costs have decreased and this problem seems to have been kept somewhat at bay for now. Even though this solution is considered temporal, better software solutions needs to developed. Massive amounts of data are generated from platforms such as social networking sites, E-commerce and Smartphones. The data generated can be classified into semi-structured, structured and unstructured data.

Velocity refers to the actual speed at which the data is generated or produced. For instance, Google currently processes 1.2 trillion searches per day globally. Facebook produces 2.7 billion like actions/day. Variety is concerned with the different formats of the data generated by different platforms. The data generated can be of any time from structured, semi-structured or unstructured. Most unstructured

Identify applicable funding agency here. If none, delete this.

data are generated from satellites and social networking sites. Videos, pictures, documents, spreadsheets and databases are in different formats which cause the data to lose its format.

Big Data challenges

The rapid upsurge in data generated in the Big data era brings about serious challenges on the acquisition of data, the storage of data, and the analysis of the data. Outdated data management systems are mostly grounded on relational database management systems(RDBMS). The main problem with such RDBMSs is that, they can only be applied to structured data. Another problem is that RDBMs are increasingly using expensive hardware. Therefore, it is very clear that traditional RDBMSs fail to handle enormous volumes and heterogeneity of Big data. In the past decade, solutions to these problems have been proposed by researchers from different perspectives. For instance, cloud computing has recently been used to meet the requirements on infrastructure for Big data. Distributed file systems and NoSQL databases have been considered good choices for solutions regarding a more permanent storage and the management of large-scale datasets.

The following section outlines challenges encountered when developing applications for Big data.

Data representation

In the era of Big data datasets consist varying levels of heterogeneity in semantics, granularity, accessibility and structure. The main aim of data representation is to ensure that is more meaningful for computer interpretation as well as computer analysis. This implies that an improper representation will diminish the value of the information or data. Furthermore, improper data representation can also hinder effective data analysis. The data structure, type and class and all the integrated technologies which enable efficient operations on datasets is reflected by an effective data representation.

Redundancy reduction and data compression

The sharply increasing data in the big data era bring about high levels of redundancy on the datasets. Data compression and redundancy reduction are two effective approaches used to lessen the unintended cost of the entire system on a premise that the possible values of the data are not affected. For instance, sensor networks generated data is highly redundant. To solve this problem the data can be filtered and compressed.

Data life cycle management

Since data is generated at unprecedented rates, current storage systems are unable to support or withstand such enormous data. This is due to slow advances in the systems for storage, computing and sensing. Most values hidden in Big data depend on the freshness of the information. Therefore, there is a need for an analytical system which will determine which data should be stored and which data should be discarded.

Analytical mechanism

The Big data analytical system should be designed to process enormous amounts of heterogenous information or data in a reasonable amount of time or within a time period which is acceptable. However, current RDBMSs cannot be expanded and are not scalable and therefore are unable to meet the performance requirements. Non-relational databases have certain advantages as opposed to typical RDBMSs in processing unstructured data and have dominated the mainstream of data analysis. Even though non-relational databases have shown improvements in performance as opposed to traditional RDBMSs, non-relational databases still have problems with regard to their performance therefore more research is still needed on the in-memory databases which will have a significant improvement in performance.

Data confidentiality

Service providers of Big data are unable to effectively maintain and analyse these huge amounts of data due to their limited capacity. They depend on professionals and tools to analyse the data which results in an increase in safety risks. For example, information about transactions usually consists of a set of complete operating data which can be used to drive key business processes. The information may be made up of sensitive information such as credit card numbers. Therefore, it has been deemed necessary to employ a third party who will analyse Big data of an organization when proper precautions have been taken to protect such sensitive information.

Energy Management

In order to process this enormous amount of data or information, mainframe computers have been used. The problem with mainframe computers is that they consume a lot of energy and therefore have drawn .so much attention from both environmental and economical perspective. As a result of the rapid increase in volumes of data generated from different sources, the processing, storage and transmission of data will consume a lot of electric energy. To lessen this, power consumption control mechanisms at a system level has to be in place for Big data.

Expandability and Scalability

Analytical systems must be designed in such a way that they support present as well as future datasets. The algorithms for Big data analysis must be designed to process present as well as future expanding complex datasets.

Cooperation

Big data analysis is achieved by combining experts from various fields who cooperate to find useful patterns on the information. A Big data network architecture must be put in place to help scientists and engineers in different fields to access different kinds of data.

Mining Big Data

Data mining refers to a process of extracting hidden predictive information from data, which helps companies to focus on important information in their data warehouse. Data mining tools are used for predicting future behaviours and trends to enable businesses or organizations to make knowledge driven and practical decisions. The contribution of data mining is manly done through a database search to evaluate hidden patterns, predictive information which may be missed by experts, due to the information being outside their expectations. The rapid upsurge in data generated in the Big data era will mean that extracting the required patterns hidden on this data will become very tedious if conventional approaches are used. In addition to that, conventional methods are time consuming and less efficient hence parallel processing of data has gained major importance.

Some data mining techniques include Anomaly detection, Association rule learning, Clustering, Regression, Classification and Summarization. This paper is mainly concerned with clustering which is a process of organizing objects in groups whose members contain some kind of similarity. Traditional clustering techniques cannot cope with this huge amount of data because of its complexity. Therefore, the main goal is to scale up and speed up clustering algorithms with minimum sacrifices to the clustering quality.

The following section outlines different types of data mining techniques.

Classification

Classification which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Credit risk applications and fraud detection are particularly well suited to this type of analysis. This technique frequently uses decision tree or neural-based classification algorithms. The data classification stage involves leaning and classification.

Types of classification methods:

Classification by decision tree induction Bayesian Classification Neural Networks Support Vector Machines (SVM)

Clustering

Clustering is a technique of identifying similar classes of objects. The use of clustering algorithms allows for the identification of dense and space regions in the object space. Therefore, the distribution pattern and the correlation between data attributes can be determined. The classification technique can also be used like clustering for distinguishing groups but it becomes very costly hence clustering is used as a pre-processing technique for attribute subset selection and classification.

Types of clustering methods

Partitioning methods Density based methods Hierarchical methods Grid-Based methods Model based algorithms

Prediction

Regression analysis models the relationship between one or more independent variables and dependent variables. Independent variables in data mining are all attributes which are already known and dependent variables are attributes to be predicted.

Types of regression methods

Linear Regression Multivariate Linear Regression Nonlinear Regression Multivariate Nonlinear Regression

Association rule

Association and correlation are used to obtain frequent item set findings among enormous data sets. This technique helps decision makers in businesses to make wise business decisions. Association rule algorithms should be able to generate rules that have a confidence level that is less that one. However, the number of possible association rules for a certain dataset tend to be very large. In addition to that, a large amount of the generated rules has little or no value at all.

Types of association rule

Multilevel association rule Multidimensional association rule Quantitative association rule

Neural networks

Neural network refers to a set of connected input/output units and each connection has a weight to represent it. Neural networks have a strong ability to derive meaning from complicated or imprecise data. It is also used to detect trends and for the extraction of patterns that are very complex to be noticed by humans. Neural networks are also well suited for prediction or forecasting needs.

Types of neural networks

Back Propagation

K-Nearest Neighbours

The K-Nearest Neighbours technique is a simple machine learning algorithm used for predicting a certain response value for a give observation using responses of all the observations in that observations local neighbourhood. This technique can also be used categorical response for classification and with a continuous response for prediction.

Regression

Regression is a data mining technique that predicts a certain number. Income, distance, weight and temperature can all be predicted using regression techniques. For instance, given the age, weight and other attributes of a child, a regression model can be used to predict the childs height. The dataset might contain attributes such as age, weight, family history and other attributes. The height of a child would be the target, and the other attributes would be predictors. During the training phase, a regression algorithm is used to approximate the actual value of the target as a function of all the predictors in the dataset. All the relationships between the predictors and target are abridged in a model. This model can then be applied to another dataset with target values unknown.

Types of regression methods

Multivariate linear regression Nonlinear regression Multivariate nonlinear regression

Summarization

It provides a more compact representation of the data set, including visualization and report generation. Data summarization provides a capacity to give data consumers a generalized view of enormous amounts of data.

Anomaly detection This is the identification of the unusual records or data errors. In other words, anomaly detection identifies unusual patterns that do not conform to the expected behaviour, which are called outliers. One of the applications of anomaly detection in business include intrusion detection, fraud detection and fault detection.

Cluster Analysis

Cluster analysis is one of the techniques used for mining big data to extract useful patterns. The goal of data clustering is to determine the intrinsic grouping in a set of unlabelled data. Traditional clustering techniques cannot cope with this huge amount of data because of their complexity. However, largescale data clustering has become a challenging task because of the large amount of information that emerges from technological progress in many areas, including finance and business informatic. Accordingly, researchers have dealt with parallel clustering algorithms using parallel programming models to address this issue. MapReduce is one of the most famous frameworks, and it has attracted great attention because of its flexibility, ease of programming, and fault tolerance.

Categories of analysis

Hierarchical methods

This approach involves creating various partitions and then evaluate the using a specific criterion. For instance, suppose there is a dataset of n objects and the partitioning methods creates k partitions of the given dataset. Each partition represents a cluster with kn. Hence, data will be classified into k number of groups which should satisfy the following requirements. i. Each partition should consist of at least one object ii. Each object is allowed to belong to a single group

Portioning methods

The technique involves creating hierarchical decomposition of a set of object or data. Hierarchical methods are classified based on how the hierarchical composition is formed. This can be achieved in two approaches i. Divisive approach This technique is also called the top-down approach. This technique starts with all the data objects in a single cluster. The cluster is then split up into smaller clusters in a continuous iteration. This is done until only one object is left on a cluster or the termination condition holds

Agglomerative approach

This technique is also called bottom-up approach. This approach starts with one data object making up a single group. Group that are close to each other are then merged in a continuous iteration. This will be done until all the groups are merged or the termination condition holds.

Density-based algorithms

This approach is based on the density and complexity functions. This technique works by growing a given cluster as long as the neighbourhoods density exceeds some given threshold.

Grid-based methods

This technique is based a multiple-level granularity structure. This means that all objects that are together will form a grid. The object space is quantized into a finite number of cells which form a grid structure. This approach has an advantage of having faster processing time.

Model-based Methods

This approach forms clusters by clustering the density function. For each of the clusters, a model is hypothesized and the idea I to find the best fit for that specific model. With this approach, the number of clusters can be deduced automatically based on standard statistics, taking noise and outliers into consideration. Hence, this approach yields very robust clustering methods.

Constraint-based method

This approach works by incorporation of application-oriented or user constraints. Constraint in this case refers to the properties of the users desired clustering or user expectation. These constraints can be specified by the application requirement or the user. Constraints are also useful for providing us with an interactive way of communicating with the entire clustering process.

Clustering algorithms applicable for MapReduce

K-Means algorithm

K-Means is considered on the most widely used clustering algorithms in clustering. This algorithm takes an input n, which refers to the number clusters that the algorithm should create and k which a set of objects. The first step of the algorithm involves randomly choosing the cluster centres and then centroids are calculated using distance functions like Manhattan distance or Euclidian distance. These steps will be performed iteratively until clusters are created.

K-Means can be implemented using MapReduce Framework as follows.

Map function

Input data is stored by the HDFS as a sequence of ¡key, value¿ pairs with each ¡key, value¿ pair representing a record. Data is distributed across all the mappers after splitting. Cluster centre information is stored on a global variant called centres. The closest centre for each record is then computed by the mapper.

Combine function

The combiner is responsible for combining intermediate data of the same mapper. This intermediate data will be located in a local disk of the host machine. Value points are also summed up in this phase. Reduce function All the output of the combine function is given to this function as input. At this stage, a summation all the values from all the nodes is produced and the results will be obtained. Thereafter, new centres are generated for further iterations. NBBBBBBBBBBB PICTURE OF MAPREDUCE BASED K-MEANS

DBSCAN (Density Based Spatial Clustering of Applications with Noise) Algorithm

This is a density-based clustering algorithm. The basic idea behind this algorithm is that points that form a dense region, are grouped together to form a cluster. In order to find this dense region, a fixed threshold value is used. DBSCAN is used mostly in research. The main advantage of this algorithm lies in the ability to handle noise data efficiently. However, even though this algorithm is more efficient than most clustering algorithms, the algorithm cannot handle large amounts of data. In addition to that, it suffers from bottleneck which is the main reason MapReduce framework is applied to this algorithm.

The MapReduce approach of DBSCAN works as follows.

Firstly, the dataset has to be partitioned and distributed among the nodes for processing. A global index structure is required by DBSCAN but it creates an extra communication cost. Therefore, a distributed index is used for efficient parallel processing. Data points belonging to the same cluster are allowed to be located on different nodes but these scattered points have to be merged. During the first phase which is called Partition with reduced boundary points. In this phase, boundary points are minimized and the input data is divided among the different nodes on the cluster. The main benefit of minimizing boundary points, is that it increases the efficiency of clustering and it also improves merging. The second phase of this MapReduce based approach is called the DBSCAN-Map. On this phase, the DBSCAN-Map algorithm is executed on every node on the cluster. KD-tree space index is used on this execution. The algorithm is executed on assigned and divided data locally. The third phase is referred to as DBSCAN-Reduce. Point indexes between partitions are found on this phase. In addition to that, the clustering ID of the points is calculated on this phase. The fourth phase is known as Merge Result. This phase involves getting results from the DBSCAN-Reduce phase and merging them. Based on boundary points obtained after merging, the global structure of clusters is established. The final phase of the algorithm is known as Relabel. On this phase, the results from local clustering on each data partition are relabelled and then global clusters are identified.

Canopy clustering

This algorithm falls under the category of unsupervised clustering algorithms. The algorithm is mainly used as a pre-clustering method which means the output produced by this algorithm is given as input to another algorithm. Pre-clustering improves the efficiency of major clustering algorithms when processing enormous datasets. The algorithm has six steps which are outlined below.

I. List containing data points which are named X is formed. II. Two threshold values T1 and T2 are selected where T1¿T2 III. A data point is selected randomly, representing a canopy centroid IV. Distance d from the centroid is calculated for all points V. If distance d is greater than threshold value T1(d¿T1), that point should be added to the canopy. Otherwise, if distance d is less than threshold value T2(d¡T2), remove the point from the list X. VI. Repeat the steps III to V until all the points have been visited.

MinHash clustering

MinHash clustering belongs to the category of Locality Sensitive Hashing(LSH) algorithms. It is more suited in situations where clustering should be done based on various dimensions of data points. This approach is a type probabilistic method. The overlap between a set of items is directly proportional to the probability of a data point being assigned to a single cluster.

Challenges in cluster analysis

The number of clusters

If the number of class labels is not known beforehand, the identification of clusters becomes a difficult task. The number of clusters have to be analysed carefully to produce the correct results otherwise, similar tuples may be broken into many or heterogonous tuples may merge which could be catastrophic if the approach used is hierarchical. This is because in a hierarchical approach, if tuples get wrongly merged, that action cannot be undone.

Identification of a distance measure

Identification of distance measure for categorical attributes is more difficult than identifying a distance measure for numeric attributes. Numerical attributes use standard equations like Manhattan, Maximum distance measure and Euclidian as distance measures. All the standard equations are a special case of MinKowski distance.

Lack of class labels

With real datasets, data should be distributed in such a way that class labels can be located easily.

Structure of the database

Real life data does not always contain clusters that are identifiable. In addition to that, the results produced by an algorithm may be affected by the order in which the tuples are arranged if the distance measure used is not perfect. With data which does not have a structure, identification of an appropriate number of clusters will not produce good results. For instance, missing values can exist for tuples, variables and also randomly in tuples and attributes.

Types of attributes in the database

Databases may not contain distinctively categorical or numerical attributes. They can also contain other types like binary, ordinal, nominal etc. In order to make calculations simpler, these attributes have to converted to categorical.

Choosing initial clusters

When utilizing the partitional approach, most algorithms require k initial clusters to be randomly chosen. If the initial clusters are not chosen properly, after a few iterations of the algorithm, clusters may be left empty.

Cluster analysis for big data

Apache Hadoop and MapReduce

Apache Hadoop is an open source software framework which uses the MapReduce software framework for distributed processing and storage of big data. The framework is made up of computer clusters which are built from commodity hardware. The core of Apache Hadoop consists of two parts, namely, the processing part and the storage part. The processing part is called the MapReduce model while the storage part is known as the Hadoop Distributed File System (HDFS). Hadoop works by splitting files into big blocks which are then distributed across different nodes in a cluster. Packaged code is then transferred into nodes to process the data in parallel. This approach uses the principle of data locality to process to process the dataset more faster and efficiently. The principle of data locality allows nodes in a cluster to only manipulate data they have access to.

MapReduce programming

MapReduce processes data points over a distributed environment which consists of commodity machines. This occurs in two phases, namely, the Mapping phase and Reduce phase. Mapping is defined by the Mapping function. The input data are split into map functions and then computations are performed which results intermediate results being produced. The intermediate results are in form of key/value pairs. After the intermediate results are generated, the data is then shuffled into corresponding Reducer to perform the reduce tasks. The Reduce phase is defined by the Reduce function which takes in a single key and processes the specified function on its associated values at a time. Overall, the data is processed through the following steps as outlined in figure 1.

Map function Responsible for reading data points as input. It then generates intermediate key/value pairs. The key represents the group number of values and the value is associated with a key. Combiner function Intermediate results that have significant repetition in their own Map task are passed through the combiner function. Therefore, this function is responsible for partial reduction before intermediate key/value pairs are passed to the reducer. Partition function Responsible for performing partitioning. Reduce function The output produced by the Map tasks is the input of the Reduce phase. For each key, the Reduce task is applied once, and its values are then processed with respect to the user-defined Reduce function which is called by each reducer once for each key.

A. *Maintaining the Integrity of the Specifications*

The IEEEtran class file is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

III. Prepare Your Paper Before Styling

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections III-A–III-E below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not number text heads—LaTeX will do that for you.

A. *Abbreviations and Acronyms*

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, ac, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as "3.5-inch disk drive".
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: "Wb/m^2" or "webers per square meter", not "webers/m^2". Spell out units when they appear in text: ". . . a few henries", not ". . . a few H".
- Use a zero before decimal points: "0.25", not ".25". Use "cm^3", not "cc".)

C. Equations

Number equations consecutively. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

$$a+b=\gamma \tag{1}$$

Be sure that the symbols in your equation have been defined before or immediately following the equation. Use "(1)", not "Eq. (1)" or "equation (1)", except at the beginning of a sentence: "Equation (1) is . . ."

D. LaTeX-Specific Advice

Please use "soft" (e.g., `\eqref{Eq}`) cross references instead of "hard" references (e.g., `(1)`). That will make it possible to combine sections, add equations, or change the order of figures or citations without having to go through the file line by line.

Please don't use the `{eqnarray}` equation environment. Use `{align}` or `{IEEEeqnarray}` instead. The `{eqnarray}` environment leaves unsightly spaces around relation symbols.

Please note that the `{subequations}` environment in LaTeX will increment the main equation counter even when there are no equation numbers displayed. If you forget that, you might write an article in which the equation numbers skip from (17) to (20), causing the copy editors to wonder if you've discovered a new method of counting.

BibTeX does not work by magic. It doesn't get the bibliographic data from thin air but from .bib files. If you use BibTeX to produce a bibliography you must send the .bib files.

LaTeX can't read your mind. If you assign the same label to a subsubsection and a table, you might find that Table I has been cross referenced as Table IV-B3.

LaTeX does not have precognitive abilities. If you put a `\label` command before the command that updates the counter it's supposed to be using, the label will pick up the last counter to be cross referenced instead. In particular, a `\label` command should not go before the caption of a figure or a table.

Do not use `\nonumber` inside the `{array}` environment. It will not stop equation numbers inside `{array}` (there won't be any anyway) and it might stop a wanted equation number in the surrounding equation.

E. Some Common Mistakes

- The word "data" is plural, not singular.
- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o".
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an "inset", not an "insert". The word alternatively is preferred to the word "alternately" (unless you really mean something that alternates).
- Do not use the word "essentially" to mean "approximately" or "effectively".
- In your paper title, if the words "that uses" can accurately replace the word "using", capitalize the "u"; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones "affect" and "effect", "complement" and "compliment", "discreet" and "discrete", "principal" and "principle".
- Do not confuse "imply" and "infer".
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the "et" in the Latin abbreviation "et al.".
- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example".

An excellent style manual for science writers is [7].

F. Authors and Affiliations

The class file is designed for, but not limited to, six authors. A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

G. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced.

H. Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 1", even at the beginning of a sentence.

TABLE I
TABLE TYPE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a] Sample of a Table footnote.

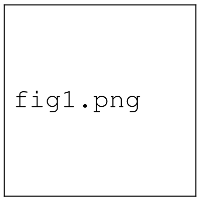

Fig. 1. Example of a figure caption.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

ACKNOWLEDGMENT

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.