INSTRUCTIONS TO REFEREES

For a contribution to be acceptable for publication in a journal, it must comprise novel material not previously published in a technical journal. The novelty will usually lie in original results, methods, observations, concepts, or applications, but may also reside in syntheses of/or new insights into previously reported research. In a regular paper, the title, abstract, introduction, and summary should be sufficiently informative to make the contributions of the paper clear to the broadest possible audience, and to place them in context with the related work. In addition to these fundamental requirements, acceptance for publication depends on a number of important criteria relating to reader interest, technical content, and presentation. To assist the referee in addressing these criteria, the Review form includes a short answer OVERVIEW (Section III) as well as an open form for DETAILED COMMENTS (Section IV). The principal intent of the Overview is to raise the kind of questions that should be addressed in assessing the paper. In other words, the Overview provides a list of the criteria referred to above and, in this sense, serves as a part of these instructions. In addition, the short answers to these questions provide a uniform synopsis of the review for both the editor and the author(s).

The essential part of the evaluation, however, is the information contained in the reviewer's Detailed Comments (Section IV). Please try to provide one or more pages of comments in this section. At minimum, please provide one half-page of comments. It is hoped that these comments will be guided by the responses indicated in the Overview, with emphasis placed on points that substantiate the recommendation to the editor. A recommendation to accept for publication, whether with no changes or with minor revisions, should be reserved for manuscripts that describe novel work and satisfy the readership, content, and presentation criteria indicated in the Overview.

If major revisions are recommended, the referee should point these out as specifically as possible and should differentiate changes regarded as optional from those judged as mandatory. If the revisions required are extensive, it is perhaps best to reject the paper and recommend preparation of a "new", heavily revised manuscript for resubmission to the journal. If a paper is rejected mainly on the basis of reader interest, the reviewer may wish to recommend submission to a more appropriate journal or conference. Papers with little or no salvageable material should be rejected outright and discouraged from later submission.

ABOUT THE JOURNAL:

This journal’s ambition is to stimulate the exchange of ideas and interaction between these two related fields of interest: Data Engineering and Knowledge Engineering. The journal reaches a world-wide audience of researchers, designers, managers and users. The major aim of the journal is to identify, investigate and analyze the underlying principles in the design and effective use of these systems. The journal achieves this aim by publishing original research results, technical advances and news items concerning data engineering, knowledge engineering, and the interface of these two fields.

The journal covers the following topics:
2. Architectures of database, expert, or knowledge-based systems: New architectures for database / knowledge base / expert systems, design and implementation techniques, languages and user interfaces, distributed architectures.
3. Construction of data/knowledge bases: Data / knowledge base design methodologies and tools, data/knowledge acquisition methods, integrity/security/maintenance issues.
4. Applications, case studies, and management issues: Data administration issues, knowledge engineering practice, office and engineering applications.
5. Tools for specifying and developing Data and Knowledge Bases using tools based on Linguistics or Human Machine Interface principles.
I. SUMMARY AND RECOMMENDATION (TO BE WITHHELD FROM AUTHOR)

Summary of Evaluation
[ ] Excellent
[ ] Good
[x] Fair
[ ] Poor

Recommendation
[ ] Accept without changes
[ ] Accept if certain minor changes are made (see Section IV)
[x] Author should prepare a major revision (see Section IV) for another round of review
[ ] Reject

If the paper is rejected, the author(s) should
[ ] Prepare a major revision and resubmit it as a new paper
[ ] Submit it to another journal or conference
[ ] Regard it as not publishable

Contribution: (Please put marks in all columns!)

<table>
<thead>
<tr>
<th>THEORY</th>
<th>APPLICATIONS</th>
<th>OVERALL QUALITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Significant</td>
<td>x</td>
<td>Excellent</td>
</tr>
<tr>
<td>Tutorial</td>
<td>x</td>
<td>Good</td>
</tr>
<tr>
<td>Possible</td>
<td>x</td>
<td>Fair</td>
</tr>
<tr>
<td>Minor</td>
<td>None</td>
<td>Poor</td>
</tr>
</tbody>
</table>

II. COMMENTS TO BE WITHHELD FROM AUTHOR

III. OVERVIEW

A) Reader Interest

1. Is the paper of current interest to a reasonable segment of JOURNAL readership?
   [x] Yes
   [ ] Perhaps
   [ ] No

2. Relative to the current level of reader interest in the paper, how is this interest likely to change during the next five years?
   [ ] Growing interest
   [x] Relatively little change
   [ ] Diminishing interest

3. Within its particular field of specialization, is the topic of the paper considered important?
   [x] Yes, definitely
   [ ] Moderately so
   [ ] Not really
B) Content

1. Is the paper technically sound?
   [ ] Yes
   [ ] Appears to be, but didn't check completely
   [x] Only partially
   [ ] No

2. How would you describe the technical depth of the paper?
   [ ] Expert level
   [ ] Appropriate for someone working in the field
   [x] Suitable for the non-specialist
   [ ] Superficial

3. Does the paper make a tangible contribution to the state-of-the-art in it's field?
   [ ] Yes, definitely
   [x] To a limited extent
   [ ] No

4. Is the bibliography adequate?
   [x] Yes
   [ ] Yes, after certain additions and/or deletions are made (see Section IV)
   [ ] No

5. To what extent is material in the paper likely to be used by other researchers and practitioners?
   [ ] Large
   [ ] Average
   [x] Small

C) Presentation

1. Is the abstract an appropriate digest of the work presented?
   [x] Yes
   [ ] No

2. Does the introduction clearly state the background and motivation in terms understandable to the nonspecialist?
   [x] Yes
   [ ] Probably
   [ ] No

3. How would you rate the overall organization of the paper?
   [ ] Satisfactory
   [x] Could be improved
   [ ] Poor

4. Relative to its technical content, is the length of the paper appropriate?
   [ ] Yes
   [ ] No, it should be lengthened
   [x] No, it should be shortened

5. Is the English satisfactory?
   [x] Yes
   [ ] No
IV. DETAILED COMMENTS (TO BE RETURNED TO AUTHORS) (Please try to provide two or more pages of comments in this section. At minimum, please provide two pages of comments. THIS SECTION MUST BE SUBMITTED ELECTRONICALLY ONE HOUR BEFORE THE DEADLINE, THE EMAIL NEEDS TO BE SENT TO ANGRYK at CS.MONTANA.EDU)

• Identify your view of the major contributions (or potential thereof) of the paper.

• Specify the major reason(s) for acceptance/resubmit/reject.

• Itemize specific revisions needed/suggested.

Detailed Comments

Major Positives

• gIndex is considered to be a benchmark graph-based indexing method. Since Freq-CI provides a better average time (more than twice) to retrieve frequent sub-graphs from a large database of graphs, thus the algorithm holds some promise. The reason for having a better retrieval time is Freq-CI would store a sequence as it is in its index. But in case of gIndex, it would store discriminative and intermediate nodes, since gIndex relies on the downward closure property. Thus if the sub sequence is one that belongs to the intermediate node, and since it does not have any id list (not a discriminative fragment) thus finding it will involve a higher time complexity.

• Since, creating the index is an offline activity, thus this approach of indexing all the sub-sequences of a graph will result in better query performance. For a relatively small database, the data retrieval times will be much faster than the gIndex, because it can retrieve the data without any calculation.

• Moreover the use of the concept of support count is pretty unique. If the support of a sub-sequence is high, they infer it to be frequent, which will
be the case in most of the cases. Again the frequent closed sequence set is more likely to exclude sequences with lower support where they follow the downward closure property, i.e., if a superset of the sub-sequence with the same support exists.

- **Lucidity of Language:** The authors have used a very lucid language to describe their research. The description of the algorithm is fairly comprehensive for a non-technical person.

- **Adherence to the format:** The author seems to have done a decent job of formatting their work and the use of reference has been fairly exhaustive. But it would have been better, if instead of using "[5] [6]" the authors would have used [5 – 6].

**Major Drawbacks**

- One of the major drawbacks of this research is its use of all the nodes for indexing. Although that will work well in fairly small databases (as the authors have used for their experiment), but, in huge databases the index creation will require a huge amount of time. Moreover the index size will grow exponentially with the size of the database. If the user prefers to have a very low min_sup (minimum support) for a fairly large database, it would be virtually impossible for FreqCI-Index to handle those kinds of situations. Thus the authors’ claim that their indexing scales better than gIndex, does not seem to hold much ground.

- Moreover, if the index is huge, then utilization and maintainance of the index seems to be a bit tough. Thus, according to my opinion for a fairly large database with a relatively small minimum support, Freq-Cl’s performance gain (in terms of data retrieval and prediction in an online environment) may not be as projected in the research. This is not a claim, but a simple observation which has to be substantiated by experimental data.

- Moreover having id_lists in a large database will have impact on the performance of the index. If the algorithm is to find out the maximum frequent fragment using the id_list, it will surely take a longer time when the id_list for each of the tuples in the index becomes huge. And in a fairly large graph database, this sequence is expected to be huge. Moreover since the research is using the downward closure properly the probability of id_list getting bigger is even more.

- The authors have not put a lot in their paper regarding searching the index for data retrieval and its related time complexity.

- The maintenance of FreqCI-Index will have a huge overhead involved in it.

- Moreover, according to me, since the algorithm is storing all the nodes in a frequent sub graph, thus insertion and deletion will be a huge overhead in this index. Although the authors have not mentioned a lot in their research regarding the time complexity involved of insertions and deletions, the best approach to follow in this algorithm according to me should be an incremental update. This will lead to the development of a single database scan algorithm for index construction.
• The lack of mathematical proofs and formal definitions in the paper decreases the technical depth of the paper. Moreover the absence of definitions of a few important component of this project is a major drawback. The readers would have to dig deep by themselves to find the definition of the important components of this research.

• The authors in Section IV have coined terms like “The support count for the new sequence ab is quite high”. Without a formal definition of high and low, it would be really hard for the algorithm to decide on which sequence is to be accepted as a frequent closed sequence and which sequence to be discarded as a result of the downward closure property.

Revisions

• As discussed above, the authors should provide mathematical proofs and definitions for the important components of their research.
• They should discuss about the time complexity of their index search and how it changes with the increase in the index size.
• Although the paper contains an experiment on the insertion time required in FreqCI-Index, but they have not elaborated on the time complexities of insertion and deletion in their index.
• They should provide separate algorithms for each of the major tasks like (searching, insertion, etc.) in their paper.
• The formal definitions of high support and low support should be duly mentioned in the paper.
• The paper must clearly state the assumptions used by the authors. One of the assumptions can be if the support for a sub-sequence is “high”, it should be considered as a “frequent sequence”.
• Although the graphs used figures 4 & 5 are 3D, but the values in these graphs are pretty hard to comprehend with naked eye. The authors should use a simpler graph or use some other means to communicate the graph to the readers.
• Since the authors have used logarithmic scales, thus they should use appropriate graphs to present it.
• There are a few repetitions (for example the abstract and the introduction starts with the exact same lines), which should be duly removed.