COMP90050, Winter Semester, 2021
Project Description

Due Date: **July 28, 5pm** Melbourne Time, for all components of project

1 Introduction

This **project forms 40% of your final mark**. The project is about creating **a survey on a recent topic in Databases**. This is a group project and each group will be 4 students in size and will be created by the lecturing team right at the beginning of the semester. Your tutors will inform you about your group members and engage with you throughout the project. Everyone in a group should contribute equally to all aspects of the project including the report and the presentation components.

The project has a **presentation and a report** that are marked separately as mentioned below but your first order of business is to choose a topic as a team from the list of topics given to start the project. There is no advantage of choosing one topic to the other. Just a discussion among team members about their liking of an area is all that is needed to choose a topic. Then, **by the end of week 1 of the semester, you need to email to your tutor about your topic of choice from the list**. (One email per group please.)

For project logistics, it is important that everyone in a group attends to the **same tutorial throughout the semester** with their teammates and this has to be the tutorial of official enrolment for our tracking purposes.

Writing a survey is a cornerstone activity that we would like you to learn as well as teamwork. In our sector, this is an activity you would need to do regularly to keep up-to-date with developments and in many cases present to your supervisors/companies that you work for.

For your topic of choice, you should initially start reading Wikipedia, newspaper/magazine, and similar articles/webpages to get a high-level idea for what the topic is about. You can do this even on day 1 of the semester and for any or all of the proposed topics to get an overall idea. These articles you check are not adequate for a survey though but good for a start.

After this initial phase, you should use **scholar.google.com or similar scholarly search engines** for performing a more detailed background search and do further reading. You should distribute the work evenly among team members. These specialized search engines give you the papers you will cite for your survey that will really count. Some sample papers per topic are given below as well to help you get started in case you feel that you are struggling with the search engines in question here. But these papers we list do not exhaustively covers the topic and should be taken as indicative works in the related area only. (It is important to note that if you login to our library with your student credentials, you will be able to access papers and some books that are returned by these engines for free in many cases.)

You are encouraged to find other survey papers that already exist in your topic as well. You do not need to start your survey from scratch basically. There would be many surveys out there potentially. Find one that is most recent. Better, find many and you will see authors look at similar but not the same set of algorithms/papers. They may also have different classifications and organization of things. These should give you an idea on what common/popular methods exist and what key comparison parameters you can have between solutions. They are also good examples on how to write surveys. You cannot use other surveys or other papers at large to copy things directly into your own surveys! These should not be the sole source of your work anyway. After you refer to these, then you need to go to individual key papers mentioned in the surveys for example and start reviewing them and form your own opinions i.e., you need to make your own judgements and categorizations basically and also check more recent works that may not appear in other surveys (although many of the categories you create would likely be similar to most survey papers of the topic/area.)

As expected from any survey, **you are expected to not only list top papers in an area one after the other, but also categorize these developments/approaches and compare/critique them**. This is at the core of a survey. A list of papers with comments attached is called an "annotated bibliography" and is not a survey and is not the purpose of this project.

It is important to note that we do not expect you to learn every paper in detail and to be perfectly comprehensive about the topic to cover all the papers. But rather cover the key papers and classifications/parameters. Digest the key directions and ideas. The number of citations a paper gets in scholar.google.com for example is an indicator about its leadership in the field, i.e., in addition to the fact that it is mentioned in other surveys.

The stages of your project can be summarized as: Background search/reading selected papers (should not take more than a week, i.e. week 2 of this semester at the latest); then organization of your report with titles and subtitles, figure captions, etc, comes… while populating sections with key bullet points/issues to mention in these sections… Finally, finishing your report by writing the details of each point and drawing figures/tables. You should prepare your presentation in tandem with your report that pretty much presents the key points in the report.

After this exercise students are expected to have a good idea in the topic and be able to answer questions during their presentation. The presentation will be done as a team at the end of the semester based on a schedule announced by the lecturer closer to the dates. The structure of the presentation should follow the report as well, i.e., in terms of the key sections it involves. The presentations will be allocated 30 mins per group including questions and setup time etc. Thus, we recommend no more than 20 slides for presentations. All members of a team should contribute to the presentation. (Note: rehearsal of presentations is the key for a successful presentation, especially when many presenters are involved.)

Following **report section-headings/structure needs to be followed (with recommended page counts per section mentioned and that aspect could vary to a degree)**:

- Identification info for students/title/abstract/etc (1 page cover basically)
- Introduction to the Topic Area (2 pages)
- Related Work Details (papers covered explained in brief in some structured manner) (7 pages)
- Comparison of Key Approaches/Papers (benefits and disadvantages from various aspects) (2 pages)
- Conclusions/Discussions and Future Directions (2 pages)
- References (1 page)

When reading the papers, please note that a technical paper is not read like a novel, i.e., not read from cover to cover sequentially, slowly, but is read in a manner that you can quickly grasp the key ideas, benefits/disadvantages. At the end of this project, in your own topics you should be able to get the idea of a new paper that you see in about 30 minutes at most! (At implementation time only, technical papers could be read to the very detail or even one can contact authors for implementations. We do not need that level of reading per method/paper from you for your project/survey.)

2 Project Administration

The deadline of this project is specified at the start of this document. **Late submissions will get a penalty of 10% per day**. The report must be **submitted as a PDF file** on LMS via a link which will be made available closer to the deadline. Handwritten reports are not accepted in any form. The report should be submitted by only one person in your group but all students' ids should be visible in the cover page.

The report should be in **A4 size paper in 11-point Times New Roman** for the main text with **1.5 line spacing** with **1-inch margins**. It should **be single column**. The report **should not exceed 15 pages (but also no less than 10 pages)**. Also, in these 15 pages putting many figures one after the other is not an acceptable report and text that describes methods is important to understand the algorithms (3500 words is adequate to cover a survey topic). Figures and Table(s) (e.g., for comparing methods) is highly recommended and crucial in some cases.

All explanations should be your own words and proper citations should be used when needed. Teams should work independently of other teams and plagiarism as usual will be checked by markers and our systems. Not sharing information about papers you found is also important as finding papers is a part of the project. Submissions are checked in our systems with other submissions, including other subjects, offerings in different semesters, websites, papers in search engines, etc through a professional plagiarism checking system.

We are not too prescriptive about your presentation format except that we highly recommend no more than 20 slides per presentation as stated above as the main constraint there is the presentation time. Having said this, presentations have to be done in electronic form over a Zoom session with all members present. The **presentation file will also need to be saved to PDF and submitted at the project deadline**. Please do not refer to external videos or other similar content and make sure presentations are standalone documents themselves. Focus on delivering good content in a clean way rather than creating "too many colourful images/videos/etc." Also, be mindful of project files you are submitting in terms of their size. Many many MBs of colourful images may later get you in trouble over a networked environment, and for the topics we cover in this subject, they are really not needed for a good report or presentation. Simple figures and tables will do much better in our view!

3 Topics

Here we give a list of **topics that you can choose from**. No other topics are acceptable. The papers listed under each topic can help you know more about related subareas/papers, and you can use them and especially their references as keys to open the door of a larger literature for each topic. You can use search engines mentioned earlier to find who references these papers in recent times as well. For these example papers, feel free to include or exclude them in your group presentation and final report at the end. Topics that you can choose from are:

- Top-k Queries in Databases
- Similarity Queries in Databases
- Solid State Drives and Databases

Note: Top conferences that you can find papers from on these topics in Databases are SIGMOD, VLDB, ICDE, SIGKDD, SIGIR, SIGSPATIAL, ICDM among others. There are also top journals we can recommend, TKDE, VLDBJ, ACM TODS. All of the publications that appear in these should pop up in scholar.google.com. Other top tier publication venues are also acceptable. There are numerous conference and journal rankings you can check for venue rankings in computer science. Starting reading papers from these venues would likely to expedite your progress.

Top-k Queries:

Yu, Albert, Pankaj K. Agarwal, and Jun Yang. "Processing a large number of continuous preference top-k queries." Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. ACM, 2012.

Lu, Jiaheng, et al. "Optimal top-k generation of attribute combinations based on ranked lists." Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. ACM, 2012.

Fraternali, Piero, Davide Martinenghi, and Marco Tagliasacchi. "Top-k bounded diversification." Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. ACM, 2012.

Mouratidis, Kyriakos, and HweeHwa Pang. "Computing immutable regions for subspace top-k queries." Proceedings of the 39th international conference on Very Large Data Bases. VLDB Endowment, 2012.

Ranu, Sayan, and Ambuj K. Singh. "Answering top-k queries over a mixture of attractive and repulsive dimensions." Proceedings of the VLDB Endowment 5.3 (2011): 169-180.

Similarity Queries:

Silva, Yasin N., et al. "Similarity queries: their conceptual evaluation, transformations, and processing." The VLDB Journal—The International Journal on Very Large Data Bases 22.3 (2013): 395-420.

Cohen, Sara. "Indexing for subtree similarity-search using edit distance." Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. ACM, 2013.

Li, Guoliang, Dong Deng, and Jianhua Feng. "A partition-based method for string similarity joins with edit-distance constraints." ACM Transactions on Database Systems (TODS) 38.2 (2013): 9.

Wang, Ye, Ahmed Metwally, and Srinivasan Parthasarathy. "Scalable all-pairs similarity search in metric spaces." Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2013.

Solid State Drives and Databases:

Do, Jaeyoung, et al. "Turbocharging DBMS buffer pool using SSDs." Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data. ACM, 2011.

Lee, Sang-Won, Bongki Moon, and Chanik Park. "Advances in flash memory SSD technology for enterprise database applications." Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data. ACM, 2009.

Roh, Hongchan, et al. "B+-tree index optimization by exploiting internal parallelism of flash-based solid state drives." Proceedings of the VLDB Endowment 5.4 (2011): 286-297

Canim , Mustafa, et al. "SSD bufferpool extensions for database systems." Proceedings of the VLDB Endowment 3.1-2 (2010): 1435-1446.

Sadoghi, Mohammad, et al. "Making updates disk-I/O friendly using SSDs." Proceedings of the VLDB Endowment 6.11 (2013): 997-1008.

4 Marking

The presentation constitutes 15% of your final marks. The **students in the same group will receive the same mark** for the presentation. Each presentation will be marked in three aspects:

- Knowledge: understanding of topic and comprehensiveness of discussion
- Delivery: clarity and engagement
- Teamwork: time management, flow, and distribution of workload

Each aspect will be **marked from 0 to 5**. The mark of your presentation will be calculated as the sum of the three aspects. Detailed marking criteria follows:

	0	**…**	**5**
Knowledge	• Does not cover any representative work related to the topic • Content is focused on a wrong topic		• Covers the representative publications/products related to the topic • Compares different approaches (methods/algorithms/products/etc.) • Covers the importance/potential of the existing work
Delivery	• Difficult to understand, e.g., full of technical jargons • Content is totally unorganized		• Slides and oral presentation are easy to understand for general audience • Presentation is clear and captures the audience
Teamwork	• Unbalanced presentation time between team members • Content from different team members is unrelated with each other • Time length of the whole presentation exceeds the limit		• Workload is well balanced between team members • Content from different team members tells a full story with fluid presentation flow • Time length of the whole presentation is well controlled • Can handle questions from the audience

The final report is 25% of your final marks. The marking is similar to presentation marking above in many ways.

	0	**…**	**5**
Knowledge Coverage	• Report is focused on a wrong topic • Does not cover any representative work related to the topic	…	• Shows a comprehensive survey of the work related to the topic • Presents representative works in the body of the report and in the references
Related Work Structure	• Papers are covered in a totally unorganized way • Development of the area is not visible to the reader	…	• Papers are well organized • Reader can see how the area has developed and/or covered by papers in subareas
Writing	• Contains many grammatical errors • Writing is difficult to understand • Writing style is not academic, e.g., using an informal tone	…	• Grammatically sound • Readers with minimal knowledge of the topic can understand the content • Uses academic writing style • Shows figures/diagrams that help readers understand the content
Format	• Format is not consistent with what is prescribed • Layout is awkward • Style of references is not consistent and academic across the reference section	…	• Format is consistent • Uses correct layout • All the references use the same proper style
Critical Analysis & Comparisons	• Plainly lists all the approaches without analysis	…	• Shows motivation, practical use and/or potential of the approaches • Categorises the approaches • Compares different approaches • Analysis from multiple perspectives

… End of Project Description..!