The Search Question Swamp, or how you can constructively develop your information need

Many of the suggested search questions contain interesting topics, but these topics also have some defects. These defects can for example be caused by language issues, methodology or the scope of the topic. In this document I discuss some general examples which may help you as a student on this course.

The rationale is that you should be able to work on your own with your search question and develop your information need. You shall not sit and wait for assistance from your lab instructor during lab 1.

Language issues

Well-written scientific texts are characterised by a precise, neutral and rather boring use of words and language. Communication of facts and information is of highest priority.

Likewise should the topic of the search question be precise formulated. The course memo states that the information need should be not too wide, not too narrow. This statement can be discussed in further detail.

Too narrow information needs are partly characterised by the search question being a ”yes-or-no”-question, or that the needs may be satisfied by one or a few documents. These needs are rarely interesting to use for the search assignment. Sometimes these needs cause the problem that you in your searches consider the relevance of material only from the specific choice of words by one single author, and not from the scientific content in the document retrieved. The information need is then identified with the presence of a few terms chosen by you in advance, when you actually should consider the information searching as any normal scientific problem. That is, you need to consider the search as a dynamical process where you change and re-formulate your search question and adapt your information need in light of the results obtained in the first searching steps and some of the relevant documents retrieved during these steps.

Exactly this dynamical process to *iterate your searches* – to constructively develop your search strings in light of previous search results and a modified information need – is something students of this course find difficult.[1] The problems with iterated searching may however be partly solved by you if you study the course literature while you try to specify you information need in writing, rather than keeping your need only in your thoughts.

[1] As a side note: *iterated searching is not the same substitution of search terms*. Iterated searching is expanding your search, *achieved partly by* finding and *adding* new search terms to your previous search strings. Simple substitution lacks the element of planning, control and precision that is required of your search.

Too wide information needs are a bit more difficult to characterize. One example is the wide information need which in fact is a *vague* information need. The vagueness may be caused by different things; one example is lack of knowledge of scientific terminology and a sloppy use of language.

Example: “technology” is a vague word. Do not say “technology” if you mean “welding”, “separation” or “gear boxes”. Do not say “programming” when you mean “implementation on a specified hardware solution”, do not say “implementation” when you mean “statistical analysis on a stratified selection of data”, do not say “certain proteins” when you mean “essential amino acids”. Sometimes you may think that the difference is irrelevant, but the difference for a reader may be semantically crucial, and methodologically this may be of importance when you start searching. I expand on this further below.

Too wide information needs may also arise if you choose to formulate a search question which contains many different *aspects.* One example can be environmental questions, life cycle analyses, carbon dioxide problems, etc. If you for example want you consider different (“green”) vehicles effect on nature, then this question contain (too) many and (too) wide aspects. What different kind of vehicles? Even if you narrow down to only cars/automobiles, then you have a very wide selection of cars on the market today (and in the research pipeline) and the search question is more suited to use in a long-term research project than in a 1,5 credit course in information searching.

So we need to narrow down our topic and define the limits or our study objects. One recent topic was “green” or “environmental-friendly” cars or rather (to be more specific! ☺), in order to avoid problems with definitions, different types of hybrid cars. But the class of hybrid cars is big and heterogenous. One limit may then be to consider the relative advantages between different kinds of hybrid solutions, electrical- versus gas- versus ethanolhybrids, just as an example.

In this way you can define one reasonable limit concerning the *objects of your study* (=what types of vehicles do you study?), but this does not make your search question sufficiently specific, since the *area of your study* (=what “relative advantages” do you consider?) is too wide.

Relative advantages, what do you mean? What type of environmental impact, which kind of life cycle analysis, what (time or geographical) frames for carbon dioxide emission do you consider? If you only consider a financial calculation on an individual level where you consider the outcome when substituting a petrol-driven car by a hybrid then your question is trivial to answer. But if you choose to consider a calculation on a group or aggregated level, where you also consider thing like carbon emission during manufacturing etc, while relating it to what parts of the global class of cars that will be replaced, then your information need is very difficult to grasp. You therefore continuously need to re-evaluate your search question and analyse what different aspects it contains.

Methodology

The last section ended with the observation that it is important to identify the essential aspects of your search question. This section contains a discussion on how you can methodically proceed when

you shall identify important terms or topics within the different aspects of your search question and how you then can generate relevant keywords from your important terms.

Suppose that you are interested in path-finding algorithms within graph theory. More precisely, you need to study how to implement path-finding in weighted graphs within some class of computer games. It follows that you cannot only consider theoretical results, but you also need to consider practical implementation and heuristic performance, while having in mind the hardware used for implementation and possibly for what different game applications the implementation will be used.

If you then choose to formulate your search question as: ”Which algorithms are optimal for path-finding in trees?” will result in leading me, your devoted author of this short pamphlet, astray. I interpret the search question as a mathematical-logical complexity theory problem, which given that the study objects are well-defined (=what class of trees are considered?), then there is a known (=already proven mathematically) solution to the question, and then the question is more suited to use as an exercise in a course in theoretical computer science than as an information searching problem.

It follows: you need to describe your search question in more detail. If you choose to re-formulate the discussion and question above, then you will approximately write something like: “What efficient path-finding algorithms exist in computer game implementation?” as your search question. This needs to be accompanied by an explanation from you that it is in particular specific heuristic algorithms on a sub-class of the class of weighted graphs (trees) that are considered as object of study, with a description of the hardware or game application issues that may be of interest.

The aspects of this search question that can be identified are roughly: “(computer) games and (path-finding) algorithms (in graph theory)”. Are these terms suitable to use as keywords?[2] That depends on database, so you cannot answer my question with certainty! What you can say is that is not sufficient to ONLY use these terms as your ONLY keywords, so you must now start to generate synonyms.

It would be a mistake here if you now choose to use “methods”, “implementations” as synonyms for “algorithms”. If your objects of study are path-finding algorithms, then you study a special class of search algorithms for graphs and trees. These are algorithms that shall be implemented in practice. Then it is a good idea if you consider algorithms that have been studied empirically, and introductory information[3] can be easily googled by you.

[2] I refrain from discussing the obvious fact that you do not use the whole complete quoted text string as our search term or use the word-by-word search question stated as our search term, but that you pick the terms that you think to be the best for the particular database that you are searching, where you have considered what type of database it is, how it indexes documents and that you use the thesaurus in the database if there is one, etc. I tacitly assume, dear Reader, that you are also a devoted reader of the course material concerning this and that you frequently consult the help pages that are available to every one of the databases used for information searching. If not, then you are participating in the wrong course.

[3] Jupp, I am talking of Wikipedia ☺: http://en.wikipedia.org/wiki/Tree_traversal

So as synonyms to the term ”path-finding algorithm” you wisely choose among terms such as alfa[4]-beta-pruning, A*, B* or why not Dijkstra´s, Kruskal´s or Prim´s algorithm (to name a few that have been named after actual human beings).

In a similar manner you deal with the term “(computer) games”. Reasonable synonyms to it are probably not ”game theory” (because then you will end up with a lot of papers in the social sciences, since they seem to be a bit fond of using game theory for research), ”electronic games”, etc. You will probably benefit from identifying synonyms by identifying what particular types of computer games that you choose to study.

You must not forget about possible hardware issues also. Perhaps is the efficiency of path-finding algorithms highly dependent on the hardware used (I do not know! ☺), are you going to implement the algorithm for a mobile unit or do you develop something to be used for a high-performing multi-core processor?

You may also never forget that the ”best synonyms” (=most productive, constructive synonyms) in one database may not be the most productive or best synonyms in the next database! How good a synonym or search keyword is to use in a database depends on the particular *indexing*[5] of the documents in that database. In a subject-specific database with thesaurus you can choose to limit your use of synonyms by mainly using the terms found in the thesaurus, since they are used by the database producer for high-quality manual indexing of the documents. In a broader/more general database covering multiple scientific fields you need to work harder or longer with generating more synonyms, writing longer and more developed search strings in order to make your search sufficiently precise. This you also need to do due to the fact that larger databases frequently apply so-called *automatic semantic algorithms* (also known as automatic term mappings, *ATM´s*, examples being lemmatization, auto-stemming, etc) and these ATM´s tend to make it harder to do precise iterative searching in the database with the help of Boolean operators, etc.

That the choice of productive synonyms is partly dependent on the database indexing is something I have seen during my years of teaching this course. I assume that you, dear Reader, is an engineering student. This means that you have taken a course in basic linear algebra and can understand the following. When you are searching in a database you do not search directly in the documents of the database (even if it is a full-text database that you are searching in), but you are searching in a “representation” of the documents. You can think of a representation of one document as a *vector*. So when you search in a database of scientific documents you search in a structured collection of tens or hundreds of millions (normally rather *sparse*) vectors. In a database with a thesaurus you can therefore choose to search very precise by only searching in the bibliographic field for the thesaurus indexing. This means (to paint you a picture), that if you choose to limit your search terms to be only used in the thesaurus field then you only traverse/search the vectors of the database in one specific position of the vectors (say position #17 for each of every of the millions of vectors in the database). Needless to say, the existence of a database thesaurus then makes it easier for you to do a precise information search in that database.

---

[4] I hate to use greek letters in Word…but I love footnotes ☺, since I am a fan of the great author David Foster Wallace.

[5] The topic of indexing and information searching and how this is done in for example large search engines such as Google is the topic of my course LI116N “Theory and models of information retrieval”.