



Lecture 9 Evaluation Methods

Chang Liu



北京大學



Assignment 2

- The purpose of this assignment is to give you an opportunity to apply the following important concepts we learned in class: Information interaction in context, search tasks, evaluation criteria, and measurement of search experience. First, choose two domain-specific IR systems (NOTE: this **excludes general Web search engines such as Baidu, Google, Yahoo!, Bing, etc.**) and write a critique of these systems. You should choose two search systems that can be used within similar contexts.

submission deadline : 2020/4/20



北京大學



Assignment 2

- Imagine that as a search expert, you are invited to contribute an article to an IT magazine. You will be expected to write a compelling, creative, interesting, persuasive, and thoughtful piece. Be explicit about your IR interaction context –Specify in what context(s) you would use such search systems. Think of at two or more search tasks including simple one and complex one that would be appropriate for these systems. Run your searches and critique the two systems. Pay attention to not only the search results themselves, but also search experiences. How easy is it for you to understand the search results? Did you experience of feeling lost? Did you feel confident about your search process? Your analysis and critique need to encompass both system performance and search experience. Be explicit about the kinds of criteria and measures you are using to evaluate and compare the systems. Use multiple criteria (about five) beyond your gut feelings or simple preferences.



北京大學



Assignment 2

- Begin the paper with a background information section in which you first describe the significance of the domain and/or systems. Introduce each system, explaining why you selected it. You should include screenshots of each system. Then report the results of your evaluation in terms of both system performance and search experience based on the criteria. Be creative in presenting your comparison data. Use graphics or tables to provide a summary or to highlight critical points. In conclusion, make a recommendation –you can recommend just one of the systems which is superior on criteria you applied or, you can recommend both systems depending on the user background, task, or context.
- I encourage you to choose systems that are related to your term project. Although this is not required, there might be some benefit to using systems in a domain that you are potentially interested in for your final project. You can also use the same criteria and measures used in this paper for your term project.



北京大學



Assignment 2

- This paper should be approximately 4 pages (single-spaced). Grading will be based on the following criteria:
- Page length (4-6 pages, single space, Times New Roman, including figures and tables)
- Introduction of two systems: 10
- Coverage of search tasks (mixture of simple and complex tasks): 10
- Degree to which search tasks and contexts are realistic: 20
- Evaluation criteria and measures: 20
- Discussions of system performance and search experience: 20
- Suggestions for the systems: 10
- Organization, formatting, and presentation of the paper: 10



北京大學



Assignment 2

- Selection of evaluation criteria and measures:
 - Will you be able to calculate recall ?
 - If you choose precision, DCG or nDCG, you can select precision@10, nDCG@10, etc.
 - Should include subjective measures, but there must be some non-subjective measures
- In the end, you should summarize your comparison of the two systems.
- Your suggestions should be based on your evaluation results.





Evaluation methods

- 一、Research methods 基本方法
- 二、Research foundation 研究基础
- 三、Data collection methods 数据收集方法
- 四、Experimental design 实验设计
- 五、Sampling 抽样

- Kelly, D. (2009). Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1),
Chapter 4: Approaches, 25-30.
Chapter 5: Research Basics, 31-43.



北京大學



一、基本方法

- Research Goals
- Laboratory and Naturalistic Studies
- Longitudinal Studies
- Wizard of Oz Studies





Research Goals

- Exploratory Studies
- Descriptive Studies
- Explanatory Studies



北京大學



Research Goals

- Exploratory studies:
 - typically conducted when little is known about a particular phenomenon.
 - often employ a variety of research methods with the goal of learning more about a phenomenon, rather than making specific predictions
 - Research questions are typically broad and open-ended and hypotheses are uncommon.





Research Goals

- Descriptive Studies
 - Focused on documenting and describing a particular phenomenon.
 - The main purpose is to provide benchmark descriptions and classifications.
 - Can lead to a weaker form of prediction via correlation analysis
 - Cannot explain why a relationship exists between two variables.





Research Goals

- Explanatory Studies
 - examine the relationship between two or more variables with the goal of prediction and explanation
 - concerned with establishing causality
 - use more structured and focused methods
 - not all explanatory studies offer explanations





Examples

Example 1: How do people re-find information on the Web?

exploratory study

Example 2: What Web browser functionalities are currently being used during web-based information-seeking tasks?

descriptive study

Example 3: What are the differences between written and spoken queries in terms of their retrieval characteristics and performance outcomes?

Example 4: What is the relationship between query box size and query length? What is the relationship between query length and performance?

Explanatory study



Laboratory and Naturalistic Studies

- Laboratory studies
 - Most experiments take place in the laboratory
 - Good: the amount of control researchers have over the study situation..
 - One perennial criticism: too artificial, do not represent real life and have limited generalizability.





Laboratory and Naturalistic Studies

- Naturalistic Studies
 - Log-based studies
 - Good: more representative of the user's true behavior
 - Drawback: the researcher has little control over the setting, which can make it hard to make cross user comparisons.
 - Difficult to administer naturalistic studies
 - more intrusive
 - privacy





Laboratory and Naturalistic Studies

- Natural experiments
- Anick (2003) conducted live trials of an interface for query expansion.
 - at a large search engine company,
 - distribute an experimental interface to a number of users and compare its use to the standard interface.





Longitudinal Studies

- A longitudinal study
 - tracks participant behavior while using a system *over an extended period of time*, as opposed to first-time usages which are what are typically assessed in formal and informal studies.
 - allows the evaluator to observe how usage changes as the participant learns about the system and how usage varies over a wide range of information needs
 - a more realistic subjective assessment
 - capture more variation in usage and behavior





Longitudinal Studies

- Dumais et al. (2003) SIGIR 03'
 - assessed the usage patterns of a personal information search tool by 234 participants over a six-week period, using questionnaires and log file analysis.
 - split the participants into two groups, giving each a different default setting for sorting results (by Date versus by Ranking).
 - to see if participants chose, over time, to use an ordering different than the default











Screen shot of SIS interface, with the Side View

Type of Item
All | Outlook | **Files**
Web Pages

Type of File
All | Word | Excel | Text
PowerPoint | PDF | Graphics
Music | Sound and Video
Other

Path contains

Date
All | Today | Yesterday
Last 7 days | Last 30 days
Older than 30 days

Document	Date	Rank	Author
Today			
 gestalt psychology <i>As a charter member, the gestalt psychologist Max Wertheimer recognized the centrality of psychology to the Graduate Department with a world-wide reputation for excellence, focusing on empirical approaches to the study of psychology.</i>	9/22/2002 4:42 PM	890	Irving Rock
 Visual Perception <i>Visual Perception: Gestalt Laws TO SEE IS... TO THINK (S. Dalf). Gestalt psychology is a mo prior to World War I. It made important contributions to the study of visual perception and</i>	9/22/2002 4:27 PM	934	Wolfgang Köhler
Last 7 days			
 CogSci/CogEng position <i>The Cognitive Science Department of Rensselaer Polytechnic Institute anticipates one or more openings beginning in candidates who have a Ph.D. in Cognitive Science or one of its contributing disciplines (i.e., AI/Computer Science, P</i>	9/20/2002 5:24 AM	645	Tyrone Slothrop
 TOC of Perception, Volume 31, SU... <i>the Microsoft Library Table of Contents Service PERCEPTION Volume 31, SUPP, 2002 The electronic alerting se library customers for business use only. Questions? Email to service@ieonline.com. (363)</i>	9/19/2002 9:25 PM	910	articles@ieonline.c...
 rademach <i>Measuring the Perception of Visual Realism in Images Research Visual Realism Define "realistic image" as able to pass as photograph Approaches to Realism Do not tell w</i>	9/19/2002 4:32 PM	879	Tyrone Slothrop
Last 30 days			
 RE: Indexing usability studies <i>Christine, Relative to developers, specifically Yoyodyne + Users, I have a lot of data about the topics you mention L about 30 minutes of video highlights. I also have some recommendations for redesigning the Open page. Help experi</i>	9/13/2002 9:55 AM	760	Oedipa Maas
 lwCsubmission.doc <i>Paper submitted to the international journal 'Interacting with Computers' BASED ON USER'S ATTENTION</i>	9/12/2002 6:21 PM	591	First Mate Gilligan
 Proceedings4.PDF <i>Waterworth Page 1 The Illusion of Being Present: Using the Interactive Tent to Create Immersive Experiences Eva an eva.lindh.waterworth@interactiveinstitute.se Phone +46 90 185136, Fax +46 90 185137 Yinteractive Institute Tools fo</i>	9/11/2002 10:04 AM	658	johwa



北京大學



Longitudinal Studies

- Dumais et al. (2003)
- They found an interesting pattern:
 - those who start with Rank order as the default were much more likely to use Date ordering than vice versa
 - suggesting that ordering information by chronology is better than by a ranking metric when searching over users' personal data collections.
 - Such an effect might not be seen in a short laboratory study.





Longitudinal Studies

- Käkä (2005) CHI'05
 - “Findex: Search Result Categories Help Users When Document Ranking Fails”
 - invited participants to use a grouping search interface over a period of two months
 - obtained Subjective responses at two points
 - after initial usage of the system
 - after the completion of the trial period.



Findex search interface



The screenshot displays the Findex search engine interface. The address bar shows 'http://'. The search bar contains the text 'jaguar' and a 'Search' button. The top right indicates 'Found about 3 140 000 results for the query.' The left sidebar lists categories, with 'atari jaguar 9' selected. The main content area shows search results for 'jaguar'.

Categories

- All results
- jaguar cars 11
- august 9
- club 9
- jaguar panthera onca 9
- mac jaguar 9
- atari jaguar 9**
- apple 8
- largest 7
- cats 6
- information 6
- released 6
- reviews 6
- powerful 5
- virtual 5
- endangered 5

Search Results:

- Atari Jaguar FAQ**
Atari **Jaguar** FAQ. Atari Archives. ... Q. What was the Atari **Jaguar**/Jaguar64? ... (41)
<http://www.digiserve.com/eescape/showpage.phtml?page=a2>
- Jaguar Interactive II -- The Premier 24-Hour Atari Jaguar ...**
... 09:45 26/Jun/04, **Jaguar** Collector, ... 18:07 25/Jun/04, **Jaguar** Collector, ... (59)
<http://www.atarihq.com/interactive/>
- Atari Jaguar VLM**
Atari **Jaguar** VLM. Mucho thanks to Joe Britt for the pix and modification details. Atari's Virtual Light Machine (VLM), was developed ... (66)
http://www.audiovisualizers.com/toolshak/vidsynth/jag_vlm/jag_vlm.htm
- AtariAge - Atari Jaguar History**
... However, after the Summer CES that year, Atari announced that the Panther was cancelled so that they could concentrate on a new machine, the 64-bit **Jaguar**. ... (95)
<http://www.atariage.com/Jaguar/?SystemID=JAGUAR>
- Slashdot | New Atari Jaguar Game Running \$1,225 on eBay**
... New Atari **Jaguar** Game Running \$1,225 on eBay. Games. ... Bill Kendrick writes, "The long-awaited Atari **Jaguar** game Battle Sphere has finally been released. ... (100)
<http://slashdot.org/articles/00/03/02/1430232.shtml>
- Slashdot | New Atari Jaguar Game Running \$1,225 on eBay**



北京大學



Longitudinal Studies

- Findings

- The responses became more positive on most measures as time went by.
- the query logs showed that the average query length became shorter over time.
- Some participants commented that they became “lazier” for some queries, using more general terms than they otherwise would, because they anticipated that the system would organize the results for them, thus allowing them to select among refining terms.
- Observing this kind of change in user behavior over time is a very useful benefit of a longitudinal study.





Case Studies

- Case studies typically consist of the intensive study of a small number of cases.
- Case studies are particularly useful when little is known about an area and for understanding more about details that sometimes get lost when averaging over large numbers of users.



Wizard of Oz Studies and Simulations



While users believe they are interacting with a real system, in reality there are one or more researchers ‘behind the curtain’ making things work.

For example, suppose a researcher wanted to study a speech user interface for querying and interacting with an IR system. Rather than building the entire system, the researcher might first want to learn something about the range of desired communications and interactions. Users might be instructed to speak to the system while a researcher sits in another room and controls the system.

- Sa, N., & Yuan, X. (Jenny). (2019). Examining users’ partial query modification patterns in voice search. *Journal of the Association for Information Science and Technology*, <https://doi.org/10.1002/asi.24238>



北京大學



二、研究基础

- Hypotheses
- Variables and Measurement
 - Conceptualization and Operationalization
 - Direct and Indirect Observables
 - Independent, Dependent, and Confounding Variables
- Measurement Considerations
- Levels of Measurement
 - Discrete Measures
 - Continuous Measures





Hypotheses

- Hypotheses follow from research questions (or theory) and state expected relationships between the concepts identified in the questions (such concepts may be more or less definable, but they are eventually represented by variables).





Hypotheses

- There are two types of hypotheses:
alternative hypotheses and null hypotheses.
 - The *null* hypothesis: states that there is no relationship or difference.
 - An *alternative* hypothesis: the researcher's statement about the expected relationship between the concepts under study.





Hypotheses

- Hypothesis testing:
 - two statements about the relationship between the evidence we collect and our hypothesis.
 - (1) our evidence allows us to reject the null hypothesis, in which case it is shown that our hypothesis provides a better (but not the only) description of what is going on
 - (2) our evidence does not allow us to reject the null hypothesis, in which case we fail to reject the null.





Hypotheses

- System A is usable
- System A is more usable
- System A is more usable than System B.





Hypotheses

- Directional or Non-directional.
 - “System A is more usable than System B”
 - “There is a difference in usability between System A and System B”





变量的种类

- Independent variables （自变量）
 - Quasi-independent variables （准自变量）
- Dependent variables （因变量）
- Confounding variables （混淆变量）
- Moderating variables （调节变量）
- Intervening variables （中介变量）

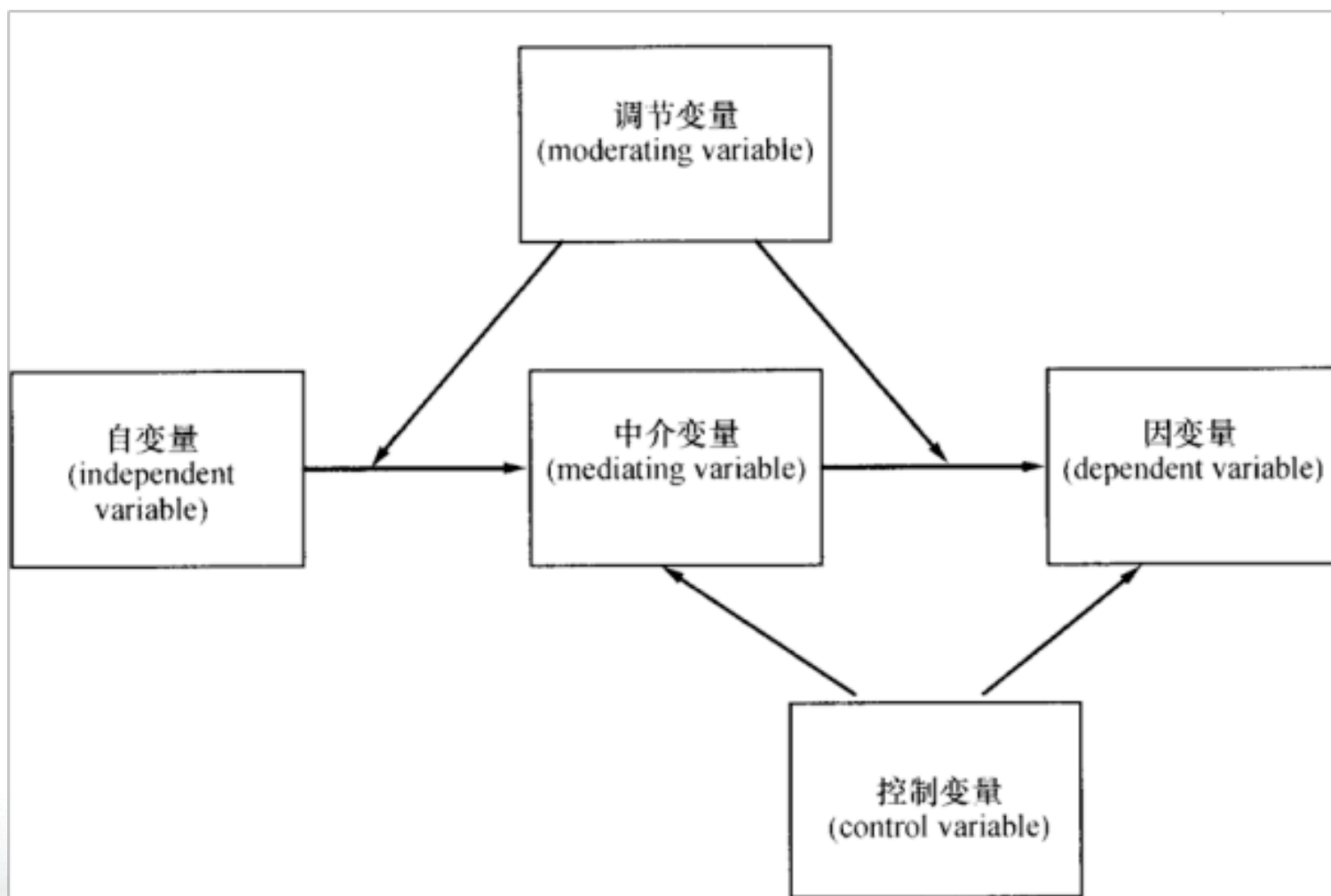




变量的种类

- Independent variables: the causes
 - Quasi-independent variables:
 - variables that can create differences in some outcome measure, but are not manipulated by the researcher.
 - Sex
- Dependent variables: the effects.







变量的种类

- Moderating variables (调节变量)
 - affect the direction or strength of the relationship between an independent and dependent variable.
 - example: consider the relationship between system, search experience and performance.

Subjects who have high search experience will perform better with the advanced system, while those with low search experience will perform better with the simple system.





变量的种类

- Confounding variables （混淆变量）：
 - variables that affect the independent or dependent variable, but have not been controlled by the researcher.
 - If a researcher realizes that such variables exist before the study starts, then the researcher can control the effects of the variables.
 - Example: compare two IIR systems, ensure that equal numbers of subjects with high and low search experience were assigned to use each of the systems.





变量的种类

- Intervening variable （中介变量）
 - provides a connection or link between an independent and dependent variable.
 - Example:
 - a larger query box → subjects to enter longer queries → better performance





三、数据收集方法

- (1) Think-Aloud
- (2) Stimulated Recall
- (3) Spontaneous and Prompted Self-Report
- (4) Observation
- (5) Logging
- (6) Questionnaires
- (7) Interviews
- (8) Evaluation of End Products





(1) Think-Aloud

- 出声思考、有声思维
- Asks subjects to articulate their thinking and decision-making as they engage in IIR.
- Inexpensive: microphone
- Problem:
 - require additional cognitive resources
 - awkward and unnatural
 - IIR task is too complex





(2) Stimulated Recall

- 有提示回忆、提示回忆
- used to collect the same type of data as think-aloud protocol, but differs in that data is collected during and after the search.
- records the screen of the computer
 - plays back the recording to the subject
 - articulate thinking and decision-making
- more expensive than think-aloud





(3) Spontaneous and Prompted Self-Report

- Subjects are not required to continuously verbalize their thoughts (as with think-aloud), but are instead asked to provide feedback at fixed intervals or when they think it is appropriate.
- get more refined feedback about the search that can be associated with particular events
- Problem: intrusive, interrupted, annoyed





(4) Observation

- In real-time:
 - The researcher is trained to focus on particular events and behaviors and takes notes that describe their observations.
- At play-back time
 - Conducted with a video camera or screen capture software
- No extra work; but may be uncomfortable
- Time-consuming and labor-intensive





(5) Logging

- A record of the user's activities and interactions
 - Capturing users' natural search behaviors outside laboratory settings
 - Run in the background without any interruption or delays
- The completeness of records varies depending on the type of logging





(5) Logging

- According to where the logging takes place
 - server: large-scale log studies conducted by search engine companies. + browser extension
 - proxy: log the communications that occur between a user and all servers to which they make requests
 - client logging: more robust and comprehensive log of the user's interactions and solves most of the problems of server-side logging





(5) Logging

- Limitation:
 - only electronic observables can be captured
- Grimes et al. (2007) “Query logs alone are not enough”
 - conducted a comparison of data collected via query log, field study and using an instrumented browser
 - the query log provided the least useful data for individual events, but the most useful for understanding the scope of user’s activities.





(6) Questionnaires

- Closed questions:
 - where a specific response set is provided (e.g., a five-point scale)
 - quantitative data
 - subjects' responses to closed-questions were significantly more positive when elicited electronically, than via pen-and-paper or interview (Kelly et al., 2009)
- Open questions
 - where subjects are able to respond in any way they see fit (e.g., what did you like most about this system?).
 - produce qualitative data





Commonly used questionnaires in IIR evaluations

- Demographic questionnaire
 - used to elicit background information about subjects. This information is typically used to characterize and describe subjects, but it can also be used to explore and test specific hypotheses.
 - usually given at the start of the study, but it can be given at the end.





Commonly used questionnaires in IIR evaluations

- Pre-task questionnaire
 - used to assess subjects' knowledge of the search task and/or topic.
 - Subjects complete this questionnaire before searching occurs so that the search experience does not bias responses.





Commonly used questionnaires in IIR evaluations

- Post-task questionnaire
 - most often used to gather feedback about the subject's experiences using a particular system to complete a particular task.
 - administered following each task.





Commonly used questionnaires in IIR evaluations

- Post-system questionnaire
 - elicits feedback from subjects about their experiences using a particular experimental system.
 - during within-subjects studies where subjects use more than one system.
 - administered after subjects finish using a system





Commonly used questionnaires in IIR evaluations

- Exit questionnaire
 - between-subjects study: similarly to the post-system questionnaire
 - within-subjects studies: elicit cross-system comparisons and ratings
 - administered at the end of the study





(7) Interviews

- A common component of many IIR study protocols.
- Deliver a set of open-ended questions
 - get more individualized responses
 - allows some flexibility with respect to probing and follow-up
- Alternatives: via print or electronic questionnaire





(7) Interviews

- Kelly et al. (2008)
- compared subjects' responses to a set of open-ended questions across three modes: *interview*, *pen-and-paper*, and *electronic*
 - subjects' responses were longer in the interview mode than in the other two modes
 - the number of unique content-bearing statements they made in each mode were about equal
 - ask more complex, abstract questions
 - simulated recall
 - Three types: structured, semistructured or open





(8) Evaluation of End Products

- Focus on work tasks, not search tasks.
- Supports a larger goal
- Approaches:
 - examination of references
 - expert assessments
 - cross-evaluation
- Used less frequently; more time consuming



(9) Concept Maps

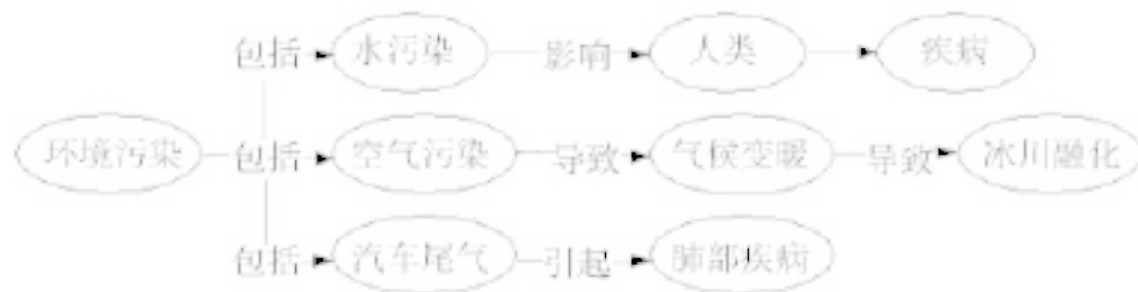


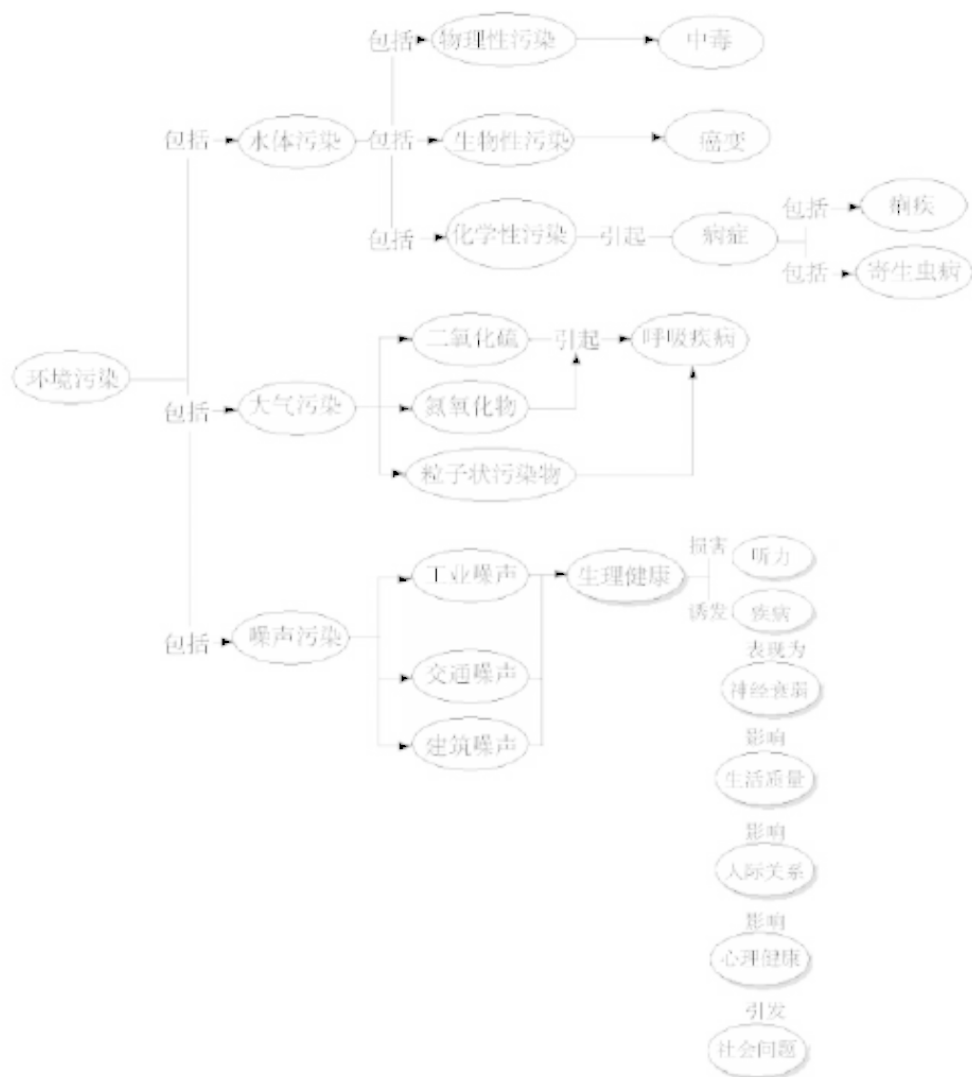
图 1 概念图示例





(9) Concept Maps





北京大学



四、实验设计

- Traditional Designs and the IIR Design
- Factorial Designs
- Between- and Within-Subjects Designs
- Rotation and Counterbalancing
- Randomization and User Choice
- Study Mode
- Protocols
- Tutorials
- Timing and Fatigue
- Pilot Testing





五、抽样

- Probability Sampling
- Non-Probability Sampling Techniques
- Subject Recruitment
- Users, Subjects, Participants and Assessors





本周的思考题 1

- 如果我们想对智能交互机器人与用户交互行为和交互效果进行评估，你能想到哪些可以研究的问题？用什么方式展开这个研究呢？



北京大学



本周思考题 2

- 这学期的所有课程都在线上授课，你觉得到目前为止，线上授课最大的优势和劣势分别是什么？



北京大學