A myth, reality and extremal principles of seeking in different search environments

Nikolai Buzikashvili

Institute of System Analysis
9 prospect 60 Let Oktyabrya, 117312 Moscow, Russia
buzik@cs.isa.ru

Abstract. Counter to commonly accepted views observations of interactive seeking behavior in different search environments show that a user behaves precisely in the same manner in traditional IR systems and on the Web. Unmediated users resort to the same tactics of successive search based on the principle of least effort (PLE). On the contrary, an intermediary in classical mediated searching follows the 'principle of the guarantied result' (PGR) rather than PLE. However, other mode of mediated search, more frequent at present time, differs from the classical one and follows PLE. A decision-theoretic model of successive search is considered; and optimal strategies of seeking are formulated for different criteria. 'Lazy user' behavior based on PLE and 'responsible intermediary' behavior based on PGR are compared with the optimal strategies. It is shown that lazy tactics usually yield suboptimal strategy under uncertainty of search environment. On the contrary, PGR-based seeking at the average is not effective.

1 Introduction

According to the widely spread point of view seeking on the Web differs greatly from seeking in 'traditional' (i.e. non-Web) IR systems. This point was mainly formed by results of the first two stages of the *Excite* project [10], [17] and by resumptive work [9]. This point is based on the fact that to represent the vague 'traditional IR systems' (TIRS) category there were selected quantitative data corresponding to the non-representative combination of environmental and non-environmental factors. However, well-known data (including those used in [9]) produce opposite evidence of efficient Web users whose information seeking behavior (ISB) is much the same as ISB in TIRS.

In Sections 2–4 we consider a myth of the Web user, give a reinterpretation of the data, and show that ISB on the Web does not differ from ISB in non-Web search environments. The factors affecting ISB are analyzed in Section 5. It is shown that today only non-environmental factors such as different subject areas may lead to differences in the quantitative characteristics of ISB.

To describe users' behavior as such and the interaction of users and search environments a decision-theoretical approach is used. In the remaining sections we consider two principles — the principle of least effort describing unmediated searching

and principle of the guarantied result describing classical intermediary's search. These are conceptual principles allowing for different ways to concretize them and possessing various technical solutions. It is shown that the actually observed ISB based on the principle of least effort is suboptimal whatever the retrieval task is and it remains the same for different combinations of environment and non-environment factors. On the contrary, classical intermediary's seeking is not effective on average and requires more terms, more steps and more time..

2 Web user myth

[7] gives a comparative analysis of ISB in three environments: the Web, TIRS and online public access catalogues (OPACs) (Table 1). This study together with observations of the *Excite* project pilot [10], first [17] and second [21] stages have brought about a commonly accepted myth about a special nature of user behavior in the Web. The results of the second stage of the *Excite* project proved to be truly mesmerizing and made the study authors see a 'dramatic tendency' witnessing "a move towards greater simplicity, including shorter queries and shorter sessions" [21].

Table 1. Comparison of seeking characteristics of the Web, TIRS and OPAC users [9]

	Web (1997/99)	TIRS (1993)	OPAC (1993)
Queries per session	1–2	7–16	2–5
Terms per query	2	6–9	1–2
% of Boolean queries	8%	37%	1%

Now the commonly suggested cause of (hypothetic) distinction of the Web ISB is not some combination of environmental and non-environmental factors but a certain feature of the Web users, who are lazy and uncurious: they make short queries (60% of queries contain not more than two words instead of 7-9 words 'prescribed' for qualified users of TIRS) and failing to find the desired result on the first two pages of the retrieved results demonstrate no willingness to go further into the remaining hundreds of pages. "Users of web search engines are a very different population" [1]. Because all non-Web users are also Web users we should suspect a mass split of personality of users into Jekylls, Dialog users and the Hydes, Web users.

3 Seeking behavior in different interactive search environments

3.1 Characteristics of seeking behavior

The commonly used characteristics of ISB fall into two groups: *quantitative* characteristics and *qualitative* ones. The first group includes the number of terms per query,

the fraction of Boolean queries and session length. The other group includes a set of seeking tactics ([2], [3], [6]), which describe successive search. The tactics are the main characteristics of ISB. If behaviors in two environments have the same quantitative characteristics but different tactics these are different behavior manners. On the contrary, any difference in quantitative characteristics with coinciding tactics doesn't allow to speak about differences in ISB in any way.

3.2 What was compared?

So far the extensive empirical data on ISB have been collected. These data fall into two quite different categories. The first category comes from external observations — transaction log analysis. The former category was formed during comprehensive studies of small groups.

Transaction logs observations are objective, representative, potentially unlimited but superficial (e.g., it remains unknown whether the user has found the needed information). In the case of comprehensive studies neither the object of search nor the users are in any way representative. A group of a dozen or a few dozens of users consists of colleagues and students of the study's authors, or, as a rule, is made up the medical students selecting bibliographies.

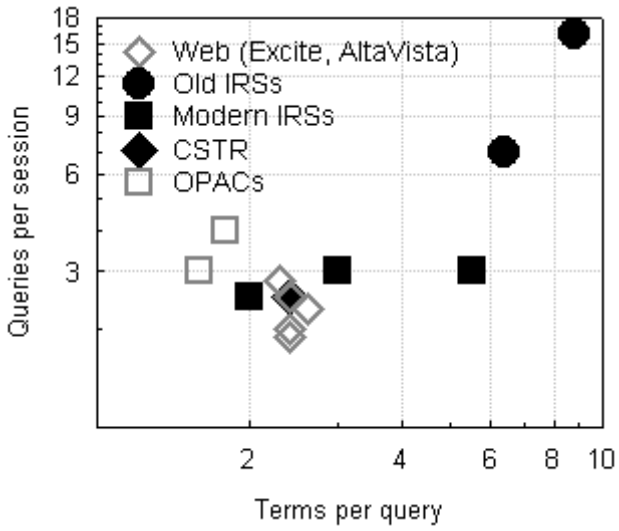

Fig. 1. Quantitative characteristics of ISB in TIRSs, in OPACs and on the Web

[9] compares ISB in three types of interactive environments: the Web, TIRS and OPACs. Is it possible to consider each of these three categories as an entity? Or is the base constituting some of the entities a common name of IR systems while differences inside the category are greater then differences between categories. Such cate-

gorization is relative and doesn't cover all search environments. The category of TIRS is enormously dispersed and is presented by fragmented and controversial data.

As seen from Figure 1, the difference between two TIRSs may be greater than the difference between modern IRSs and the Web. However, the same system may dramatically differ from its earlier version. The recent IRSs have no differences with the Web search engines except a database specialization and partly database size.

3.3 Observations

Let us consider the data on the environments whose interfaces provide more possibility for comparative analysis, i.e. the Web and TIRS. The conclusions about special character of ISB of the Web users are based on comparison of the results of the *Excite* project and the data of three studies of small groups of TIRS users. These studies were conducted in the different period. The comparisons in each position (session length, query length, etc.) are based only on one of the three studies (other two studies give no corresponding data) [9]. But studies of ISB in IR systems conducted in the recent decade (i.e. simultaneously with Web studies) reveal less quantitative differences between IBS of the Web users and IRS users.

3.3.1 Transaction log studies of the Web and IRS users

***Excite* project** ([10], [17], [18], [21]) is not the largest study (e.g. [16] reports 123 millions queries in august-september 1998 to *AltaVista*) but it is the most sophisticated and it is longitudinal study. The study was released in 1997–2001. It dealt with queries addressed to *Excite* in 1997 (pilot and first stage), in 1999 and 2001.

New Zealand Digital Library project. [11] considers queries to *CSTR* (*Computer Science technical reports*) collection of the New Zealand Digital Library. The *CSTR* base is full-text and contains 50000 documents neither categorized nor annotated, that is it is similar to the Web with the exception of its specialization.

Table 2. Data of the *Excite* project [10][17][18] and the *CSTR* (New Zealand DL) study [11]

	Excite				*CSTR*
	1997 pilot	1997 stage 1	1999 stage 2	2001 stage 3	1996/7
Topics	All				CS
Number of queries (thousands)	51	1000	1000	1000	32
Number of sessions (thousands)	18	211	326	262	16
Mean queries per session	2.8	2.5	1.9	2.3	2.5
% of modified queries	22%	52%	39.6%	44.6%	40%
Mean terms per query	2.3	2.4	2.4	2.6	2.4
% of Boolean queries	9%	5%	5%	10%	20%
Mean screens viewed per query	2.3	1.7	1.5	1.7	1–2

All users of the *CSTR* collection are specialists in the computer science, who are considered to be good searchers unlike, for example, medical students. Besides, these users are familiar with the subject area of the collection unlike, for example, librarian intermediaries assisting to medical students. Thus, we should expect the exemplary ISB greatly different from ISB of uncultivated Web users.

There were used two interfaces — Boolean and free-text. During the first stage of the study the Boolean interface was used as default and during the second stage the default interface was free-text one. In both cases 66% queries were enter in default interface, and in both cases the fraction of Boolean queries was about 20%.

As seen from Table 2, except the fraction of Boolean queries, the results of the *CSTR* study *are identical* to the results of the *Excite* project, that is quantitative characteristics of ISB of the best users of TIRS are *just the same* as characteristics of ISB of 'undeveloped' Web users.

3.3.2 Comprehensive studies of seeking behavior in TIRS

Session length. According to [9], an average session consists of 7 queries. However, [14] and [19] speak about it 2–3 or 3 queries respectively.

Query length. On the one hand, it is universally acknowledged that a number of terms in 'classical' query is a big: 7.9 for novice users and more bigger (14.4) for experienced users [5] or 8.8 for novice searcher and less (7.2) for experienced searcher [7]. These data are results of comprehensive studies of small groups conducted in 1981 and 1993. These data not only dramatically differ from observed (in 1996–2001) in the *Excite* project but also demonstrate too big variety and too different tendency to be compatible.

On the other hand, more recent results show that query consists of 1–3 terms [11], [14] or queries without synonyms consist of 3 terms and queries with synonyms — 5.5 terms in average [19].

Use of Boolean queries. 37% as fraction of Boolean queries to TIRS given in [9] as opposite to 5–10% of the Web queries may look impressive as long as it remains unknown that these are data the only study [15] where the remaining 63% of queries contained just one term. It means that Boolean interface was used and the fraction of single-term queries was twice as much as the fraction in the Web queries (!), which can hardly prove that traditional IRS are used by advanced users.

Thus, contemporary studies of ISB show no significant differences in quantitative characteristics between the Web and non-Web environments.

4 Dispelling the myth

Nobody teaches users of these systems [12], queries in these systems contain just 1 or 2 terms, the session length is 2–3 queries and users' tactics are simple. This is what says the myth about Web users. And yet these statements describe not the Web but OPAC [12] and TIRS.

The conclusions made about special character of ISB of the Web users are based on fragmental data of three TIRS studies conducted a decade before the Web studies. But as we have seen, other studies reveal less or no quantitative differences between behavior of the Web and TIRS users.

As a result, we should say about special character of ISB in TIRS. While the results of all Web studies are very closed, the TIRS studies give very dispersed results. The cause of this fact is analyzed in the next section.

As for qualitative ISB data, they indicate to the absence of differences both within the categories and between categories. [19] describes the behavior of students in TIRS Lisa unknown to them before. During the first session students used a wide range of tactics. In the last session they use only one term as an initial query, and a set of modification tactics in the last session reduced to simple *AND*- and *OR*-expansions with one term and changing or deleting one term, i.e. *perfectly to the same tactics* which are used by 'bad' Web users who made "a move towards greater simplicity, including shorter queries and shorter sessions, with *little modification (addition or deletion) of terms in subsequent queries*" [21]. All users in all studies use sets of simple tactics or move to these sets during learning to interact with a search environment.

Already in 1991 Wildemuth et al. [20] while observing medical students using TIRS showed that *the most common ISB is to use simplest tactics*. The same results were given by all studies of ISB in both the Web and TIRS without exceptions.

5 Factors of differences in seeking behavior

Although the Web user myth dispelled, the question of the importance of the different factors, which determine quantitative characteristics remains open. The following *non-environmental* factors affecting behavior should be explicitly considered:

- *user*: specific, specialized, and average, where 'specialized' means some feature, e.g. profession;
- *topics/subject area*: specific topic, specialized, and average, where 'specialized' means some subject area (medicine, computer science, entertainment) and 'average' denotes all topic categories.

To avoid additional assumptions while comparing behavior in different environments we need to have available observations corresponding to the same values of the user and area factors, i.e. we should know behavior of the same type of users searching the same type of topics. If we know that medical students who search some medical topic in TIRS behave in different manners than the same users searching the same topic on the Web we can speak about different behaviors. On the contrary, if different users searching different topics manifest different behaviors we have no reasons to state anything.

Although a lot of studies of ISB in non-Web IRS were conducted, these studies, especially earlier studies cover exactly the same combination of factors and investigate ISB of "medical students seeking full bibliography in small base", i.e. these studies consider specialized topics, specialized vocabulary and uncommon, recall-oriented task.

However, all the Web studies correspond to those combinations of non-environmental factors, which *do not intersect* with combinations presented in the non-Web studies (Table 3). We see neither coinciding nor similar combinations of non-environmental factors in the compared Web and TIRS data. The partial exception is the *CSTR* study [11]: specialized users search specialized topics (computer science) and do it in a manner similar to an average Web user seeking average topic.

Table 3. Combinations of non-environmental factors presented in ISB studies of the Web and TIRS users. (Notations: ♦ — TIRS studies, ⊕ — Web studies. Typical combinations of the factors are given without references.)

Topics / Users	Specific	Specialized (subject area)	Average (all topics)
Specific	♦	♦	⊕ ([1])
Specialized	♦	♦ ([11])	
Average			⊕

While a comparison of the Web and non-Web studies is very limited and requires additional assumption, a comparison of different non-Web studies shows that only topic specialization and a seeking task orientation (common precision-oriented tasks and uncommon recall-oriented tasks) determine quantitative characteristics of ISB.

The properties of the search environment traditionally considered as factors significantly affecting ISB are:

- *response time* of IR system;
- *interfaces/search methods* used by IR system;
- *size* of DB and a *number*/*fraction* of relevant documents.

Which factors and to what extent are significant? (1) Response time was essential only for IRSs of 1960-70th. (2) As it follows from the data of the *CSTR* study [11], Boolean and free-text interfaces cause no significant differences of quantitative characteristics. As it follows from the data compiled in [9], search methods (e.g. including Inquery) cause no significant differences of quantitative characteristics. (3) Only the number of relevant documents is a significant but threshold factor — this number should be *sufficient* rather than a huge. When this number is small, users need to expand their queries by synonyms. On the other hand, if this number is small for an average query then the base is small, which is not actual case now.

The differences in the quantitative characteristics are caused by difference in topics and by difference in sizes of bases. It is not that all Web users collect bibliography for master's thesis. The same medical student enters 3 search terms when looking for "sleep apnea syndrome" and enters 1–2 terms when searching pornosite. Searching in a sufficiently big base (the Web search engine or Medline) this student doesn't have to expand initial 3-terms query by 4-terms synonym ("sudden infant death syndrome") and construct Boolean query as *((slep AND apnea) OR (sudden AND infant AND death)) AND syndrome* while searching in a small base he has do it.

Thus, besides non-environmental factors only the DB size may lead to different quantitative characteristics of ISB in different environments.

6 Lazy searcher and responsible intermediary

6.1 Unmediated search and the principle of least effort

The task of search is not a self-valued problem but a supplementary one for a searcher. A mathematician may spend years to find the finest proof of the previously proven theorem, but nobody spends any time to construct the 'finest' query or sequence of query modifications. A searcher follows the principle of least effort (PLE) (e.g., [4]). This principle is applied at all levels of information seeking from selecting a source (e.g. a talk with a colleague, seeking in a real library or search in the base) to specific steps of query modification in successive interactive search.

PLE is realized through lazy tactics — searchers 'say' in queries less than they can say. This style is *interactive* — the current query (including the first query) is not considered as the final one and a search is a successive ('multiple-step').

At first sight lazy tactics look like the tactics of minimum risk but it is not quite so. Here we can see two excluding opposite cases. The first is when lazy tactics is reasonably careful. That is, if the user is certain that a more complex query modification won't lead to worse results and his certainty exceeds some threshold of certainty, he will choose this complex modification. If the user is not certain enough he will use either the simplest modification (expansion with one term) or modification certainty of which exceeds some threshold. Suppose results become worse. Then, if the modification at the previous step was the simplest at the next step the user has to make a new modification instead of this unsuccessful one. If the modification at the previous step was complex the searcher doesn't know which of its components cause worse results so he cannot exclude any of these components with some degree of certainty. So any part of it may be used and the previous step is useless. To cut down on the number of useless steps, searchers use complex modifications only being quite certain of their use. Thus, the lazy tactic is also the tactic of minimum risk.

The second case. Suppose the searcher has found a reference to the article he needs in any PDF-document. Then the tactic of minimum risk is to manually enter the whole article specification. In this case the rest consists of one step. On the contrary, lazy tactic means that the searcher enters only part of the specification. This tactic is less careful and at the average requires more than one step. In this case the lazy tactic differs from the tactic of minimum risk.

The first case is more frequent and the lazy, precision-oriented ISB is the most typical with one exception. This exception is a mediated search.

6.2 Mediated search and the principle of guarantied result

ISB of classical intermediaries is quite a different ISB. Namely, an intermediary *doesn't follow* the principle of least effort. S/he follows the "principle of the maximum query completeness", or *the principle of the guarantied result* (PGR): s/he tries to miss nothing, because s/he doesn't know what is important for user's information

needs. Sometimes this principle results in the same tactics as the principle of the minimum risk (as in the above mentioned example).

Table 4. Searcher in unmediated and mediated searching

Searcher	Ability to recognize pertinence	Searcher's aim	Tactic at each step
'patron'	yes	precision of the search results	least effort tactic
intermediary	no or partial	a mostly complete query	maximum completeness tactic

Interacting with the IR system a searcher faces environmental uncertainty of distribution of desired documents and that of system behavior as such. Besides these uncertainties, an intermediary faces uncertainty of users' information needs. Namely, an intermediary usually knows something about information needs of the users and the subject area. Intermediaries try to make more complete and more general queries [13]. The maximum completeness tactic is the result of the uncertainty of the information needs rather than intermediaries' features. An intermediary 'says' in queries all s/he knows. This style is *non-interactive* and corresponds to the situation when, in fact, each query (including the first query) is considered as the final query. An intermediary formulates queries in such a manner, which guarantees the presence of documents corresponding to all possible interpretations of the patron's information need in the [observable part of] search results.

However, the more confidential an intermediary is with the information need the more s/he diverts from the maximum completeness tactics. Now the most frequent manner of mediated search is not iterative user-librarian interaction but non-iterative procedure: a patron-chief formulates retrieval task and an intermediary-subordinate carries out this task. Unlike the classical mediated search, this intermediary knows the subject area and follows PLE.

Table 5. Intermediary's ISB depending on his/her ability to localize external information need

Ability to localize	A query formulated by an intermediary	Search principle	Typical situation
no	reproduction of the stated need	minimum risk [least effort]	special need (atypical)
partial	broader query	guarantied result	classical mediated search (rarely)
perfect	Narrow query	least effort	non-classical mediated search (typical)

7 Optimal seeking strategies under uncertainty and two types of searching

Our experience tells that the result of lazy doing is doing anew and that fast modifications at each step may slow the process on the whole. On the contrary, the experience tells that careful approach is the most effective one. Is it true in the case of ISB?

PLE is a universal principle [22] but it is a conceptual rather than an operational one. To elaborate an operational (formal) model we need to specify this principle in each case. Let us consider the following successive search model formulated in terms similar to the Bates-Fidel language of query modifications [2], [3], [6]. Let a searcher use any combination of any terms. Some of the combinations are 'magical', i.e. the combinations for which the viewed part of the retrieved results includes desired documents. The aim of the search is to guess one of the magical combinations. The user may compare the results of successive steps (a query modification possesses worse or improved retrieved results) and add, delete or replace term(s) depending on the changes in the results. This framework allows to consider the following criteria of *successive* search.

The criterion of minimum number of added terms (including terms of the initial query). First, the strategy of the least modification (addition, deletion, replacement by exactly one term) at each step is optimal for any degree of uncertainty. In the special case of certainty (see the above mentioned example of a reference in the PDF file) other optimal strategies also exist — to enter any number of terms at each step.

The criterion of minimum number of steps. The greater uncertainty is the less number of terms may be modified at each step. The strategy of minimum (one-term) modifications is the best for great degree of uncertainty.

The criterion of minimum summary time. The simplest measure of effort is the time used. If we suppose that a system response time and a 'user response time' (to view retrieved results) are intrinsic, we come to the criterion of minimum steps. More realistic assumption is that a user response time depends on the similarity of results of the current and previous steps (first-order dependency). In this case simple modifications became more preferable, and we come to the 'weak' minimum steps criterion.

Thus, lazy tactics usually form optimal or suboptimal behavior under uncertainty. The rules of query modification don't depend on any factors as well on a search goal (precision or recall). The number of steps of the successive search and the length of the query depend on the factors.

Contrary to PLE-based ISB, a search following to PGR (an classical intermediary's search itself) requires more terms, more steps and more time.

8 Conclusions

We have dispelled the Web user myth and shown that a user searches in the same manner in different search environments. We have excluded some factors presumably determining quantitative characteristics of seeking behavior. We have considered unmediated and mediated searches, which follow different principles — the principle of least effort and the principle of the guarantied result. We have considered some

criteria of ISB and shown that 'lazy behavior' at each step of the successive search yields suboptimal strategies. On the contrary, a classical intermediary's manner of seeking is effective only in the case of uncommon information needs.

References

1. Aula, A. Query Formulation in Web Information Search, *Proc. ICWI* (2003), 403-410
2. Bates, M. Information Search Tactics. JASIS, 30(4) (1979) 205-214
3. Bates, M. Where should the person stop and information interface start? IPM, 26(5) (1990) 575-591
4. Bates, M. Task Force Recommendation 2.3. Research and design review: Improving User Access to Library Catalog and Portal Information. Final Report. (2003)
5. Fenichel, C. Online searching: Measures that discriminate among users with different types of experience. JASIS, 32 (1981) 23-32
6. Fidel, R. Moves in Online Searching. Online Review, 9(1) (1985) 61-74
7. Hsieh-Yee, I. Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. JASIST, 44(3) (1993) 161-174
8. Holscher, C., Strube, G. Web search behavior of internet experts and newbies. Proc. of 9th WWW conference (2000) 337-346
9. Jansen, B., Pooch, U. A review of Web Searching Studies and a Framework for Future Research, JASIST, 52(3) (2001) 235-246
10. Jansen, B., Spink, A., Saracevic, T. Real Life, Real Users, and Real Needs: A Study and Analysis of User Quaries on the Web. IPM, 36(2) (2000) 207–227
11. Jones, S., Cunningham, S., McNab, R., Boddie, S. A transaction log analysis of a digital library. Int. J. on Digital Libraries, 3(2) (2000) 152-169
12. Meadow, C., Marchionini, G., Cherry, J. Speculations on the Measurement and Use of User Characteristics in Information Retrieval Experementation. Canadian J. of Information and Library Science, 19 (4) (1994) 5-7
13. Nordlie, R. Unmediated and Mediated Information Searching in the Public Library. ASIS 1996 Annual Conf. Proc., (1996) http://www.asis.org/annual-96/ElectronicProceedings
14. Seiden, P., Szymborski, K., Norelli, B. Undergraduate Students in the Digital Library: Information Seeking Behavior in an Heterogenous Environment. ACRL National Conference (1997) http://www.ala.org/acrl/paperhtm/c26.html
15. Siegfried, S., Bates, M., Wilde, D. A profile of end-user searching by humanities scholars. JASIS, 44(5) (1993) 273-291
16. Silverstein, C., Henzinger, M., Marais, M., Moricz, M. Analysis of a very large Web search engine query log. ACM SIGIR Forum, 33(3) (1999).
17. Spink, A., Wolfram, D., Jansen, B., Saracevic, T. Searching the Web: The Public and Their Queries. JASIST, 52(3) (2001) 226-234
18. Spink, A., Jansen, B., Wolfram, D., Saracevic, T. From E-Sex to E-Commerce: Web Search Changes. Computer, March 2002, 107-109
19. Vakkari, P. eCognition and changes of search terms and tactics during task performance. Proc. of RIAO 2000, 894-907
20. Wildemuth, B., Jacob, E., Fullington, A., de Blieck, R., Friedman, C. A detailed analysis of end-user search behaviors. Proc. of 54 ASIS Annual Meeting (1991) 302-312
21. Wolfram, D., Spink, A., Jansen, B., Saracevic, T. Vox Populi: The Public Searching of the Web. JASIST, 52(12) (2001) 1073–1074
22. Zipf, G. Human behavior and the principle of least effort. Addison-Wesley, Cambridge, MA (1949)