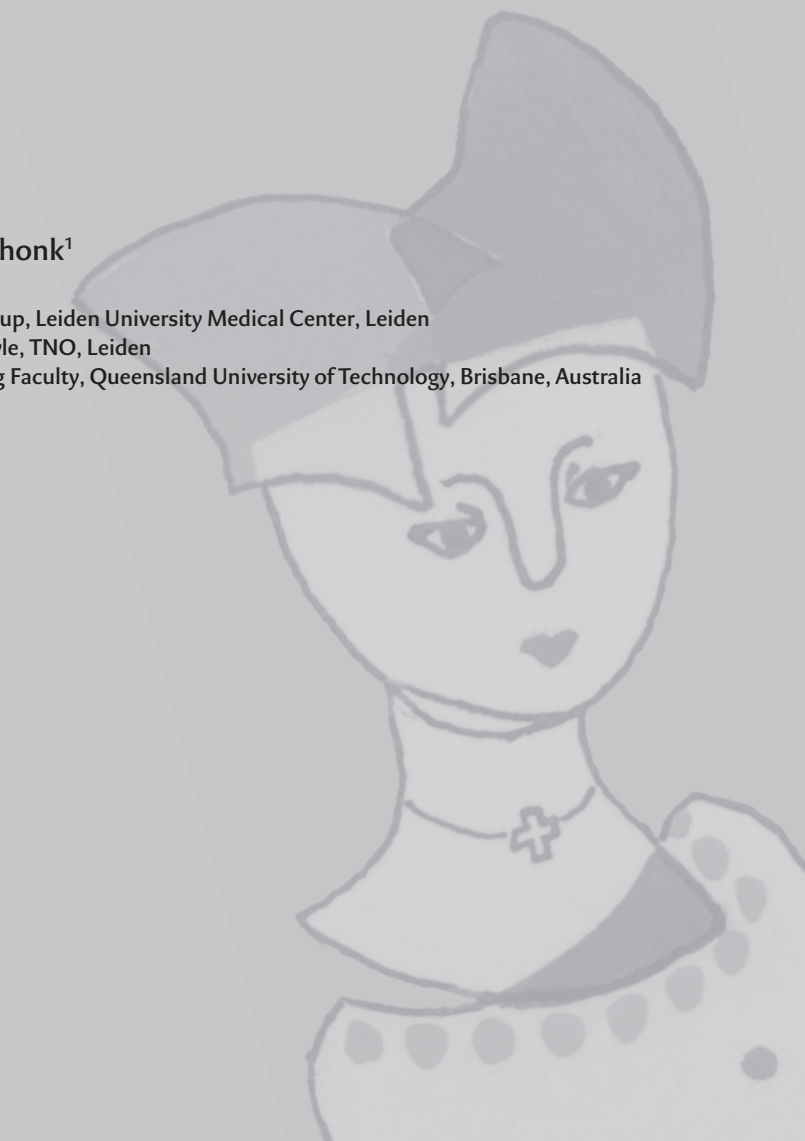
CHAPTER 7

Searching for breast cancer stories on a website using verbose natural language queries: an exploratory study

Regina Overberg[1]
Wilma Otten[2]
Andries de Man[1]
Eduard Hoenkamp[3]
Bertie Zwetsloot-Schonk[1]

[1] Clinical Informatics Group, Leiden University Medical Center, Leiden
[2] Expertise Centre Life Style, TNO, Leiden
[3] Science and Engineering Faculty, Queensland University of Technology, Brisbane, Australia



Abstract

Online patient stories provide other patients with support and information. Several search facilities have shown their usefulness in finding relevant stories, but also have limitations. Latent Semantic Indexing (LSI) may overcome these. This mathematical technique places stories and queries in a multidimensional space: distances between them provide information about similarity in content. No human reading is necessary to tag and categorize stories. Searching involves typing natural language queries, observing a list of retrieved stories (adapts while typing) and narrowing search results by refining queries (after accessing stories). We examined how twenty-four breast cancer patients use an LSI application to search for stories. Natural language queries ranged from complete sentences to one or two keywords. Sixteen participants refined their search queries. Most participants used the facilities that LSI offers and were quite satisfied with the search process and the dynamic list of retrieved stories. More research is needed to implement LSI in searching for patient stories.

1 Introduction

Stories of other patients can provide patients with emotional support, information, reassurance and practical advice [1]. Nowadays, the Internet has become an increasingly important source of patient stories [2-7]. Several websites offer patients an comprehensive set of stories of other patients (e.g.[8,9]). A crucial element in these sites is how people can retrieve the stories that fulfill their needs. This paper examines a search facility based on Latent Semantic Indexing (LSI), whose development does not require human reading to tag and categorize stories.

Regular search facilities include the following. The standard search facility of the browser (CTRL-F) can be used to search for specific words on a website. However, synonyms are not searched for and only the open page is searched through. Some major web search engines offer a facility that makes it possible to search within a specific website and allows users to type long queries in natural language. Yet, this facility also searches in other parts of the website than the part of interest and it is unclear which search algorithms are used and whether the whole query is used (trade secret).

When special search facilities are developed, they are usually based on concepts instead of text words. For example, articles in the PubMed database can be searched for using the MeSH terms that convey concepts [10, 11]. Developing such a search facility requires two steps: 1) establishing a list of concepts; and 2) tagging the content of a document to these concepts. Concept-based search facilities are also applied to patient stories, like on the healthtalkonline website [8]. Studies showed that patients who search for stories of other patients appreciate concept-based search facilities [1, 12]. Yet, to develop a concept-based search facility human effort is required (establishing concepts, tagging content) [13]. Other disadvantages are that users are restricted to pre-defined search categories [11] and that tagging content is subject to human interpretation.

Latent Semantic Indexing (LSI), a technique from the field of Information Retrieval (IR) [14], seems to overcome the previous mentioned shortcomings. A set of documents is placed in a multidimensional space: this space is a vector space that has the words in the documents as coordinate axes (the dimensions) and the documents as points in that space. In constructing this multidimensional space it is taken into account to what extent words and dimensions are distinctive. A word that occurs equally often in each document is not distinctive and will hardly count in constructing the space. Two or more dimensions that almost coincide will be reduced to one dimension [15]. The technique is purely mathematical: meanings of words and documents are not examined. The idea of LSI is that documents that are close together in the multidimensional space will be quite similar in content. Studies have shown that LSI divides a set of documents in the same clusters human subjects would assign when they are asked to group the documents by content [14, 16].

In the context of searching for illness stories LSI can be applied as follows [17]. A set of patient stories is placed in a multidimensional space. Subsequently, patients can search for stories by typing any text in a search box: one or more words, phrases or sentences. These natural language queries are considered mini-stories and are placed in the same multidimensional space as the set of stories. In this step, the whole natural language query is used. Stories with the shortest distance to a query are retrieved and

presented to the user. While typing a search query, the location of the query in the multidimensional space continuously changes and thus also the list of stories retrieved changes. Figure 1 shows that the initial query is close to Story 1, but that the final query is close to Story 5. After accessing one or more stories, a search query can be refined without having to start a new search. Research has shown that the longer the search query, the more targeted the search results [18].

To our knowledge LSI has not been studied before in the context of searching for illness stories.

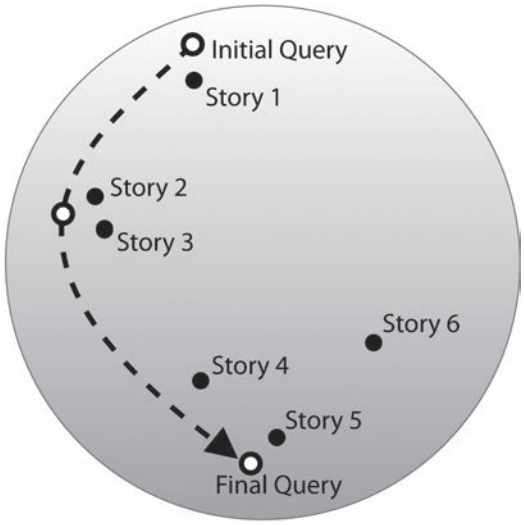


Figure 1 Example of a multidimensional space containing several stories. During the typing of the search query, it will travel through the document space [17].

In this paper we describe a study in which 24 breast cancer patients used a search facility based on LSI to search for stories of other patients. The aim of this exploratory study was to get an idea whether LSI would be suitable to disclose breast cancer stories. To this end we wanted to get insight into whether participants use the LSI search facility optimally: the search results of the LSI application are most targeted when a user types verbose natural language queries and refines search queries after accessing stories. Furthermore, we wanted examine whether satisfaction with an LSI search facility is in the same range as satisfaction with a story topics search facility or with a writer profile search facility. If breast cancer patients use the LSI search facility optimally and are quite satisfied with it, it will be more cost effective to disclose stories with LSI than to disclose stories with a story topics search facility and/or a writer profile search facility, since building an LSI search facility takes less time and less human effort than building a story topics and/or a writer profile search facility.

The research questions we addressed, were:
- How do breast cancer patients use the facilities LSI offers?
  a. Do participants refine their search queries after accessing stories in order to get more specific results?
  b. In what ways do participants use natural language in their search queries?
- How satisfied are breast cancer patients with the search process and the dynamic list of retrieved stories?

2 Methods

2.1 Design and procedure

Study design
We developed a search application based on LSI to search for breast cancer stories. The stories were downloaded from the website of *De Amazones* [19]. This website was founded by a group of young women with breast cancer and provides stories that are spontaneously submitted by patients for publication on the website. In April and May 2008 women with breast cancer were invited to the Leiden University Medical Centre to use our application in a computer room. They could not see each other's screen and one of the investigators was present for questions. The women were encouraged to use search queries consisting of natural language and to search for whatever they wanted. They could search for and read in the stories as long as they liked. Their search queries and the time involved were automatically saved in a database. At the end women completed a final questionnaire about their satisfaction with the search process and their background characteristics, also automatically saved in the database. For details see Appendix I and [20].

Recruitment process
Recruitment announcements were disseminated online via several breast cancer forums and offline via local newspapers, the magazine of the LUMC and several support groups. Women who subscribed were sent additional information, including an informed consent form. Twenty-four women returned the form. They received a book token for their participation and restitution of travel costs.

Ethical aspects
The board of *De Amazones* foundation gave us permission to conduct the study. The women who submitted their story to the website of *De Amazones*, i.e. the writers of the stories that could be searched for in the LSI application, were not identifiable since they used nicknames when submitting their story. Our research proposal was presented to the Ethical Committee of the Leiden University Medical Centre. The Committee concluded that our study involved no medical intervention and that we could proceed.

2.2 Development of the application

Set of stories
In January 2007 all 171 stories were downloaded from the website of *De Amazones* [19]. The stories were all written in the Dutch language and in the first person. All stories were 'completed' individual stories; they were not part of an interactive forum. Length, structure and content of the stories differed. The mean length of the stories was 759 words (SD=723 words) and ranged from 55 words to 5,112 words (median 568 words).

The document space
Latent Semantic Indexing [14] was used to place the stories of *De Amazones* in a multidimensional space. After removing stop words, the words in the set of stories were reduced to word stems. A word stem-by-story matrix was constructed, with columns for stories and rows for word stems. This matrix was filled with weighted values. For each cell in the matrix the following ratio was calculated: the frequency of the word stem in the story divided by the number of stories the word stem occurred in. The number of dimensions of the document space that was generated in this way, was lowered by applying a special technique purported to represent meaning underlying the stories rather than words [15]. For technical details see Appendix II.

The search queries of the participants went in-real-time through the same process as the stories and were also placed in the document space. Stories with the shortest distances to the search query were retrieved and presented to the user. A search query of one word that did not appear in the set of stories could not be positioned in the document space.

Study website
The search page of the search facility consisted of an explanation, a search box, a Reset button and a dynamic list of stories retrieved (Figure 2). The search application was explained as follows:

> *On this page you can search for stories on the website of De Amazones. You can search by typing text in the box below. This can be your own story, for example, but it can also be some text about a topic you want to know more about. While you are typing, stories that are similar to the text you type will be retrieved. The more text you type, the more accurately stories can be retrieved.*

There was no Search button, since the application started searching automatically when a participant completed or deleted a word or stopped with typing for 3 seconds.

For each story retrieved the writer's nickname was presented as well as zero to five pink ribbons (Figure 2). The number of pink ribbons indicated the degree a story matched the query. The more pink ribbons, the higher the similarity. Retrieved stories were presented in descending order of the number of pink ribbons. Ten stories were presented to the user and all stories that had the same similarity degree as the tenth story.

Retrieval of stories continued non-stop during the typing of a search query: stories could disappear from the list, other stories could appear and the number of pink ribbons could change. While a participant was typing a search query, she could see which changes her typing caused in this dynamic list.

A user could at anytime access one or more stories in the dynamic list by clicking on the nickname of a writer. A user who accessed stories could return to the search page at anytime. After accessing stories a user could choose to refine the search query already typed by adding or removing text or to completely remove the search query already typed and start typing a new query. With the Reset button text in the search box could be removed.

To finish the application participants could click on the Questionnaire button. After clicking this button they could not return to the application.

Figure 2 Screenshot of the search page showing the LSI search facility.

2.3 Data analysis

Data from the database were imported in the statistical software package SPSS 17.0. Descriptive statistics were performed. Time spent using the application was defined as the total time elapsed between starting with typing the first search query and clicking on the Questionnaire button. Time spent reading stories was defined as the time spent using the application subtracted by time spent searching. The mean reading time per story per participant was calculated by dividing the time reading stories by the number of stories that was accessed. In order to interpret the degree of satisfaction,

the results of the present study were compared to the results of an earlier conducted online randomized controlled experiment in which the same questions about search satisfaction were asked [20]. The answers to the last question of the final questionnaire asking whether participants had any remarks about the study, were examined in order to get insight into participants' remarks regarding satisfaction with the search facility.

Participants' search queries (sequence, verbatim texts) and the number of stories accessed thereupon were imported from the database into MS Excel to count the number of searches a participant performed. We defined a *new search* as typing a query in an empty search box: i.e. at the start or after complete removal of a query already typed. Furthermore, for each participant the verbatim text of the query that contained the most words was word-counted. Also, the number of stories accessed was counted. A story that was accessed multiple times in the same search session, was counted multiple times.

Moreover, refinement fractions were calculated. A *search refinement* was defined as altering the query already typed after accessing stories (i.e., adding text at the end, the beginning or in the middle of a query; or, removing and/or replacing part of the query text). A search could contain more than one refinement. A participant's *refinement fraction* was defined as the number of searches with refinements divided by the total number of searches. The refinement fraction lies between 0 and 1. The closer to 0, the smaller the percentage of searches refined by a participant. The closer to 1, the greater the percentage of searches refined by a participant.

The verbatim texts of the queries provided insight in the use of natural language in the queries. The grammatical structures of the queries (range: complete sentences – separate keywords) and the descriptions of information needs (personal experience, in what person, in question form, etcetera) were examined. The search queries that are quoted in this article as illustrations were translated from Dutch into English.

3 Results

3.1 Participant statistics

Table 1 shows the demographic and disease characteristics of the participants and their use of the Internet and *De Amazones* website before study participation. One participant was excluded from the analyses due to not having accessed any stories.

Table 1 Baseline characteristics of the participants (n=23).

Baseline characteristics	N (%)[a]
Demographic characteristics	
Age in years (mean, SD)	51.9 (9.4)
Married or living together	16 (70)
Children	19 (83)
Religious	8 (35)
Higher professional education or university degree	11 (48)
Employed	14 (61)
Disease characteristics	
Time since diagnosis in months (mean, SD)	61.6 (62.4)
Diagnosed with one tumour	15 (65)
Size of tumour ≥2 cm	16 (76)
Cancer in axillary lymph nodes at diagnosis	12 (55)
Metastases to other parts of the body	0 (0)
Breast conserving surgery	9 (39)
Mastectomy	17 (74)
Radiation therapy	13 (57)
Chemotherapy	16 (70)
Hormonal therapy	13 (57)
Cancer free	21 (91)
Use of the Internet and De Amazones website	
Daily Internet use	19 (83)
Familiar with searching online for specific information	23 (100)
Familiar with accessing fellow patients' stories on the Internet	21 (91)
Visited De Amazones website at least once before participation	15 (65)
'Rather well' or 'well' familiar with De Amazones website	10 (67[b])
Read half or more of De Amazones stories before	7 (47[b])

[a] n(%) is shown unless noted otherwise
[b] Percentages based on the n=15 participants who visited *De Amazones* website before.

3.2 Use of the LSI application

Search behaviour

Time spent using the application varied from 8.9 to 76.2 minutes, with a mean of 48.6 minutes (SD=20.0). Participants performed on average 7 searches (SD=5; range 1-21). Their longest search query contained 2 to 130 words, with a mean of 22 words (SD=30). Participants accessed on average 26 stories (SD=14; range 6-52) and their mean reading time per story was on average 2.0 minutes (SD=1.8; range 0.6-7.0).

Figure 3 and 4 provide illustrations of the search sessions of two participants with different search strategies. Figure 3 shows a participant who performed one search in which she refined her search query twice after accessing stories. She started with typing a search query (SQ1) and accessed 13 stories of the list of stories retrieved. Then, she added text to the query already typed (SQ2), accessed 6 stories of the changed list, and again added text (SQ3) and accessed 6 stories. In her whole session, she accessed 25 stories, four of which she accessed twice (these are double counted). Her longest search query consisted of 13 words (=SQ3, word count in Dutch). Her refinement fraction is 1.0 (=1 search with refinements/1 search).

Figure 4 shows a participant who performed 12 searches. In her first search she refined her search query once after accessing stories. In her whole session she accessed 45 stories, six of which she accessed twice (these are double counted) and one of which she accessed three times (this one is triple counted). In three of her searches she did not access any stories. Her longest search query consisted of 6 words (=SQ2, word count in Dutch). Her refinement fraction is 0.08 (=1 search with refinement/12 searches).

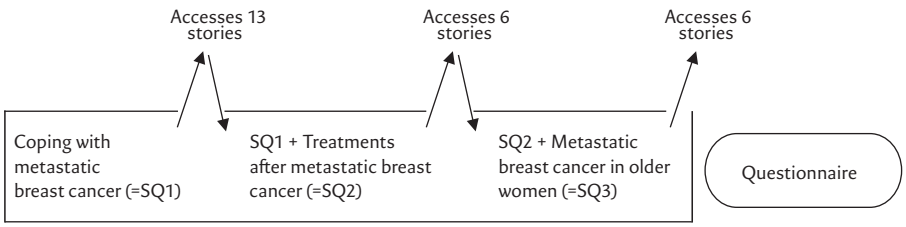


Figure 3 Search session of a 61-year-old participant who spent 76.2 min using the application with a mean reading time of 2.6 min per story. SQ=search query. SQ2 is formed by expanding SQ1, SQ3 by expanding SQ2.

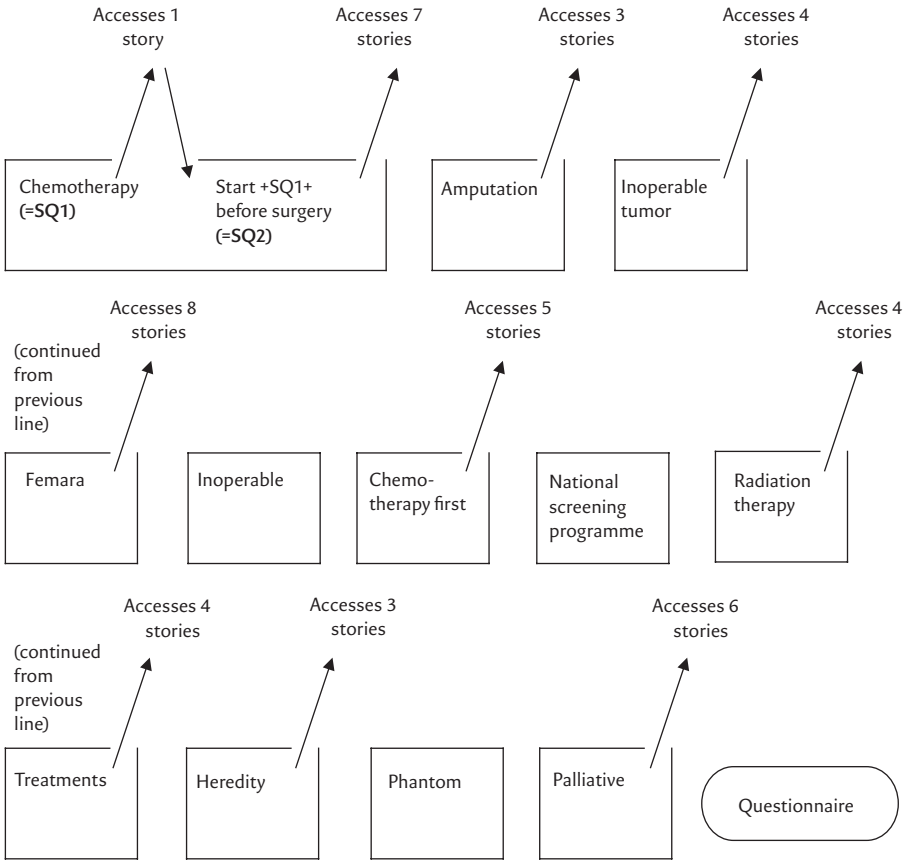


Figure 4 Search session of a 59-year-old participant who spent 58.1 min using the application with a mean reading time of 1.1 min per story. SQ=search query. SQ2 is formed by expanding SQ1.

Refinement of search queries

Figure 5 shows the refinement fractions of the participants. The mean refinement fraction of the participants was 0.25 (SD=0.30) and the median was 0.17. Seven participants (30%) refined none of their searches and therefore had a refinement fraction of 0. Two participants (9%) refined all of their searches and therefore had a refinement fraction of 1.

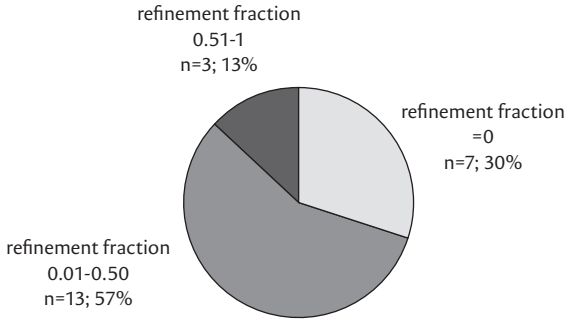


Figure 5 Refinement fractions of the participants (n=23).

Natural language use in search queries

The analysis revealed that natural language was used in three different ways (note that a participant who performed more than one search and/or refinement can use several categories in her session):

1. *Complete sentences*. Twelve participants used complete sentences in their queries. These sentences can be divided in three subcategories:
   a. *Own experience in the first person*. Four participants typed their own experience in the first person. For example:
   'After chemo and radiation therapy I have developed many problems with my condition and joints. I am therefore very limited in doing things.' (54-year-old participant)
   b. *Description of information need in the first person*. Three participants decribed what they were searching for in the first person. For example:
   'I want to know more about possible metastases.' (58-year-old participant)
   c. *Questions in general or literally directed to other patients*. Seven participants formulated questions. One participant addressed her questions directly to her fellow patients:
   'The weight gain since taking Tamoxifen, how did you lose those extra pounds? This was hard for me, and also the menopausal symptoms were disappointing.' (59-year-old participant)
   Others asked questions more in general, for example:
   'Breast reconstruction and nipple reconstruction, is that painful?' (35-year-old participant)
2. *Topics described in a short grammatical structure*. Fourteen participants formulated search queries in which they used a short grammatical structure to describe the topic they were looking for. Examples are:
   - 'experiences with breast reconstruction and lymphedema' (58-year-old participant)
   - 'coping of children after they heard that their mother has breast cancer' (38-year-old participant)

3. *Topics described in one or two keywords*. Fourteen participants used search queries consisting of one or two keywords (word count in Dutch). Examples are:
   - 'heredity' (53-year-old participant)
   - 'hormonal therapy' (38-year-old participant)

   Two of these fourteen participants used no other type of search queries: they used only search queries consisting of one or two keywords.

3.3 Satisfaction with the search process

Table 2 shows participants' satisfaction with the search process. All outcomes measured on a scale from 1 to 5 scored 3 or higher. The outcome measure 'opinion about the list of stories displayed after a search' scored highest with a mean of 4.4 (SD=0.6). The mean of the outcome measure 'overall satisfaction with the search facility' was 6.7 (SD=1.7) (scale 1-10).

Table 2 Participants' satisfaction with the search process (n=23). Higher means indicate better outcomes.

Outcome measure	Mean (SD)
Opinion about the search facility (range 1-5)	3.9 (0.7)
Opinion about the list of stories displayed after a search (range 1-5)	4.4 (0.6)
The extent to which search options enable finding information one was looking for (range 1-5)	2.9 (1.0)
Recommendation to others and future own use (range 1-5)	3.7 (0.9)
Overall satisfaction with the search facility (range 1-10)	6.7 (1.7)

The results for satisfaction with the LSI application lay in the same range as the results for satisfaction with a story topics search facility or with a writer profile search facility [20]. The story topics search facility group scored on 'overall satisfaction with the search facility' a mean of 7.3 (SD=1.4) and the writer profile search facility group scored a mean of 7.1 (SD=1.6). The outcome measure on which LSI seems to score somewhat higher is 'opinion about the list of stories displayed after a search'. On this outcome measure the story topics search facility group scored a mean of 3.6 (SD=0.9) and the writer profile search facility group a mean of 3.8 (SD=0.9).

The answers to the last question of the questionnaire asking whether participants had any comments about the study, yielded some useful remarks regarding satisfaction with the LSI search facility. Seven participants commented on the content of *De Amazones* stories. In their opinion the stories were predominantly written by young women and by women who just received the diagnosis and contained quite personal experiences. These participants were more interested in stories of older women who received the diagnosis longer ago and in stories containing experiences that were supported by scientific research. One participant commented on the search facility. According to her there was sometimes no link between her search query and the stories retrieved and when using certain keywords as search query no stories were retrieved at all. This participant used predominantly one or two keywords as search queries.

4 Discussion

The purpose of this study was to get insight into how breast cancer patients use the facilities that LSI offers in searching for stories of other patients and how satisfied they are with an LSI search facility. This LSI search application was built without the need for human tagging and categorizing of the stories and offers the user the benefits of searching in natural language queries, a dynamic list of retrieved stories which adapts while typing a query, and -after accessing stories- refining queries to narrow search results.

In building the LSI search facility we experienced that the fine-tuning of the multidimensional space took quite some time and energy. An explanation for this might be that breast cancer stories are quite similar: each of the stories discusses treatments and coping with illness, although with a slightly different focus. LSI can correctly cluster a set of texts about entirely different topics [14, 16], but may have more difficulty with texts that are more similar.

The result that participants differed in search behaviour is consistent with previous studies. Analyses of large numbers of user queries on Web search engines revealed that users differ in the number of queries and results pages viewed per query, and in the mean number of terms per query [21, 22]. Kim [23] found that a person's search behaviour –such as time spent in retrieving information- was associated with cognitive style, information search experience and information search task. So, in order to interpret the patients' search behaviour for stories, patients' information need and reasons for searching should be known. For a patient with a specific information need conducting one search might indicate a good performance, while a patient with a wider, not yet fully defined information need might need to conduct more searches for a good performance. In this exploratory study such relationships were not examined.

More than two thirds of the participants refined a search query after accessing stories, nearly one third did not. By refining a query search results may be more targeted. Explanations for not refining any of their search queries might be that participants are not used to search in this way or that the stories they had retrieved so far already satisfied their information need. In previous studies it was found that most users searched one query only and did not follow with successive queries [21, 22]. In future research it would be informative to find out the reasons why some participants did not refine their search queries.

Participants used natural language queries in three ways: 1) complete sentences, 2) short grammatical structures, and 3) keywords. Short grammatical structures and keywords were most used as queries. Our results were in line with the results of Zeng et al. [24] who found that consumers who searched for health information on the MEDLINEplus website tended to use short and general text queries (rarely more than one or two words). Spink et al. [21] showed in their study of consumer use of a large Web search engine that the mean number of terms in queries was 2.4. An explanation might be that people are most accustomed to using short queries. Their experience with other search engines might be that general, short queries retrieve results that give sufficient satisfaction. In addition, it takes less time and energy to type a short query than to type

a whole sentence. Yet, with LSI it is expected that the longer the search query, the more targeted the stories retrieved.

Seven participants addressed in complete sentences a question directly to other patients. We might say that this is in line with the intentional stance theory of Daniel Denett: they saw the LSI application as if it were a 'live' interaction partner and assigned an intention to it [25].

Participants were as satisfied with the LSI application as with two other search facilities used in a previous study, that is, a story topics search facility and/or a writer profiles search facility [20]. However, the two studies are not entirely comparable since the participants of our earlier study were younger, participated online without direct contact with the investigator, and the groups were larger [20].

A striking result is that the outcome measure 'Opinion about the list of stories displayed after a search' is relatively positive in the LSI group compared to the other two search facilities. The LSI-users might have a high satisfaction on this outcome measure because they could observe the list of retrieved stories while typing a search query and could immediately see any changes in this list. With qualitative research, for example a think-aloud study, it could be examined why participants have a high satisfaction on this outcome measure.

The Internet is highly dynamic. The focus is shifting from searching for information using search engines into sharing information via social media [26]. Members of the online patient community PatientsLikeMe, for example, send their experiences to other members who -according to their profiles- may most likely benefit from them [27]. Yet, we think that search facilities remain important for patients who do not wish to create a personal profile on social media.

4.1 Limitations

The present study was an exploratory study. The number of participants was relatively low. Also, the mean age of the participants was somewhat higher than the mean age of the writers of the stories the participants could search for (mean age at diagnosis 35.8 years (SD=6.2)) [28]. The difference in age between writers and participants may have affected participants' search satisfaction. Furthermore, the questionnaire did not go into detail on participants' reasons for their search behaviour and satisfaction.

4.2 Conclusion

Our study showed that participants did use the facilities that LSI offers in searching for other patients' stories. The majority of the participants refined their search queries and used short grammatical structures or complete sentences in their search queries. Furthermore, satisfaction with LSI was in the same range as satisfaction with a story topics search facility or a writer profile search facility. Participants appreciated the dynamic list of retrieved stories during the typing and formulating of their query. Given these promising results of the use of the LSI search facility by patients to search for stories of other patients, further research would be useful. With an extensive quantitative research the LSI search facility can be statistically compared to other search facilities. A qualitative research can provide insight into the motivations of patients.

Acknowledgements

The authors wish to thank the participants for their time and interest. We also wish to thank the organizations that disseminated the call for participation.

Funding

This work was supported by the Netherlands Organization for Scientific Research (NWO). The paper is part of the Narrator project addressing the problem of storage and accurate retrieval of illness narratives, conducted under the umbrella program ToKeN2000.

References

1. Rozmovits L, Ziebland S. What do patients with prostate or breast cancer want from an Internet site? A qualitative study of information needs. *Patient Educ Couns* 2004; 53: 57-64.
2. Ziebland S, Chapple A, Dumelow C, et al. How the internet affects patients' experience of cancer: a qualitative study. *BMJ* 2004; 328: 564-569.
3. McTavish FM, Gustafson DH, Owens BH, et al. CHESS (Comprehensive Health Enhancement Support System): an interactive computer system for women with breast cancer piloted with an underserved population. *J Ambul Care Manage* 1995; 18: 35-41.
4. Herxheimer A, McPherson A, Miller R, et al. Database of patients' experiences (DIPEx): a multi-media approach to sharing experiences and information. *Lancet* 2000; 355: 1540-1543.
5. Pitts V. Illness and Internet empowerment: writing and reading breast cancer in cyberspace. *Health (London)* 2004; 8: 33-59.
6. Hardey M. 'The story of my illness': personal accounts of illness on the Internet. *Health (London)* 2002; 6: 31-46.
7. Chou WYS, Hunt Y, Folkers A, et al. Cancer Survivorship in the Age of YouTube and Social Media: A Narrative Analysis. *J Med Internet Res* 2011; 13(1): e7.
8. Health Experience Research Group. Database of personal and patient experiences (formerly Dipex), www.healthtalkonline.org (accessed 11 May 2008).
9. BreastCancerStories.org, www.breastcancerstories.org (2006, accessed 18 December 2011).
10. PubMed. Combining MeSH Terms Using the MeSH Database, http://www.nlm.nih.gov/bsd/viewlet/mesh/combining/mesh2.html (2011, accessed 18 December 2011).
11. Richter RR, Austin TM. Using MeSH (Medical Subject Headings) to Enhance PubMed Search Strategies for Evidence-Based Practice in Physical Therapy. *Phys Ther* 2012; 92: 124-132.
12. Overberg RI, Alpay LL, Verhoef J, et al. Illness stories on the internet: what do breast cancer patients want at the end of treatment? *Psychooncology* 2007; 16: 937- 944.
13. Ziebland S, Herxheimer A. How patients' experiences contribute to decision making: illustrations from DIPEx (personal experiences of health and illness). J *Nurs Manag* 2008; 16: 433-439.
14. Deerwester S, Dumais ST, Furnas GW, et al. Indexing by Latent Semantic Analysis. *J Am Soc Inf Sci* 1990; 41: 391-407.
15. Hoenkamp E. Unitary Operators on the Document Space. *J Am Soc Inf Sci* 2003; 54: 314-320.
16. Landauer TK, Dumais ST. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychol Rev* 1997; 104: 211-240.
17. Hoenkamp E, Overberg R. Computing Latent Taxonomies from Patients' Spontaneous Self-Disclosure to Form Compatible Support Groups. *Stud Health Technol Inform* 2006; 124: 969-974.
18. Hoenkamp E, Bruza P, Song D, et al. An Effective Approach to Verbose Queries Using a Limited Dependencies Language Model. In: *Advances in Information Retrieval Theory: Second International Conference on the Theory of Information Retrieval, ICTIR 2009* (eds L Azzopardi, G Kazai, S Robertson, et al.), Cambridge, UK, 10 September-12 September 2009, pp. 116-127. Springer.
19. De Amazones, young women with breast cancer (In Dutch: De Amazones: jonge vrouwen met borstkanker), www.de-amazones.nl (accessed 2 January 2007).
20. Overberg R, Otten W, De Man A, et al. How breast cancer patients want to search for and retrieve information from stories of other patients on the internet: an online randomized controlled experiment. *J Med Internet Res* 2010; 12(1): e7.

21. Spink A, Wolfram D, Jansen BJ, et al. Searching the web: the public and their queries. *J Am Soc Inf Sci* 2001; 53: 226-234.
22. Jansen BJ, Spink A, Saracevic T. Real life, real users, and real needs: a study and analysis of user queries on the web. *Inf Process Manag* 2000; 36: 207-227.
23. Kim KS. Information seeking on the Web: Effects of user and task variables. *Libr Inf Sci Res* 2001; 23: 233-255.
24. Zeng QT, Kogan S, Plovnick RM, et al. Positive attitudes and failed queries: an exploration of the conundrums of consumer health information retrieval. *Int J Med Inform* 2004; 73: 45-55.
25. Dennett DC. True believers: The Intentional Strategy and Why It Works. In: Dennett DC (eds) *The Intentional Stance*. Cambridge: MIT Press Mass, 1987, pp. 13-36.
26. Lober WB, Flowers JL. Consumer Empowerment in Health Care amid the Internet and Social Media. *Semin Oncol Nurs* 2011; 27: 169-182.
27. Frost JH, Massagli MP. Social Uses of Personal Health Information Within PatientsLikeMe, an Online Patient Community: What Can Happen When Patients Have Access to One Another's Data. *J Med Internet Res* 2008; 10: e15.
28. Overberg R, De Man A, Wolterbeek R, et al. Spontaneously published illness stories on a website for young women with breast cancer: do writers and themes reflect the wider population? Submitted.
29. ORACLE. Dutch (nl) Default Stoplist, http://web.archive.org/web/20020111230441/downloadwest. oracle.com/otndoc/oracle9i/901_doc/text.901/a90121/astopsu4.htm (2001, accessed 29 November 2007).
30. Kraaij W, Pohlmann R. Porter's stemming algorithm for Dutch. In: *Informatiewetenschap 1994: Wetenschappelijke bijdragen aan de derde STINFON Conferentie* (eds LGM Noordman and WAM De Vroomen), Tilburg, The Netherlands, pp. 167-180. Tilburg: STINFON.