Computing Latent Taxonomies from Patients’ Spontaneous Self-Disclosure to Form Compatible Support Groups

Eduard Hoenkamp
Nijmegen Institute for Cognition and Information
&
Regina Overberg
Leiden University Medical Centre
Chance Encounters

• They happen in real life
• They happen on the internet
• And they are often the most valuable

Perhaps you can nudge the chances a bit in your direction: constrain the places you go, a conference, a bar, the internet…

…except the internet is way too big.
Where we want to go

Our focus is women diagnosed with breast cancer.

Sharing experiences may help to cope with depression, anxiety, and fear, often occurring after the diagnosis.¹

On many internet sites patients tell the story of their illness.

So if those sites could be found easily, it could help patients:

• share experiences

• form support groups


The Narrator project wants a system that supports communication among patients, by providing narratives supplied by patients.

Our take here is to collect such narratives by looking them up on the internet. But how?

We identify a patient with the narrative she writes, her diary.

The diary is taken as a very elaborate query to find information on the internet. It takes the place of the 2 or 3 keywords you would normally type into your browser.
Steps to take

- Term weighting
- Dimension reduction
- Clustering

What are these?!?
Information Retrieval in 3 slides

The *vector space model* for documents (bag of words approach).

Suppose there were only 2 words, A and B

Distance = $\cos(\alpha)$
Searching for query AB

Diagram showing the search for the query AB.
Dimension reduction
(e.g. linear regression)

This reduces from 2 to 1

For document space: Alleviates the lexicon problem.
E.g. from 10000 to 3
Interface for the study
Interface for the study (cont’d)

Type your query:
I am interested in immigration and refugees -- I learned about Canada's new program to integrate refugees -- how are newcomers motivated to enter the labor market -- what skills and education do newcomers possess like for instance a university degree.

[Options and settings...]
Moving diaries

A diary as a very elaborate, very precise query
Hierarchical clustering
Steps we took to compute a latent taxonomy

- Term weighting: straight $ti/idf$
- Dimension reduction: wavelet transform
- Clustering: cosine distance + furthest neighbor

We called the result the *latent taxonomy* (as we don’t know what the classifiers are)
Steps we took to compute a manifest taxonomy

- Careful selection of characteristics of the diaries (medical history, treatment, prognosis, feelings, interaction with family, ...)

- Singular value decomposition (PCA) + hierarchical clustering

Both taxonomies (latent and manifest) give similar results for extreme distances (very close and very remote documents)
The manifest taxonomy
Combining the taxonomies

The document space can be extended with the manifest factors, and this hyperspace is reduced again.

This requires a method to weight the added dimensions.
Future work

• Find method to weight manifest dimensions or retain mapping from Procrustus. Either can be used to classify new diaries.

• Investigate how compatibility among patients is associated with the eventual clustering. (e.g. nearest or furthest neighbor? Selecting dimensions?)
Conclusion

By taking the diary of a patient as her proxy, we can define a ‘compatibility’ space

• in which new diaries can be placed

• which define potential support groups

For women diagnosed with breast cancer, the prospect lies in building virtual support groups to complement friends and family.

Which, in the absence of those, would make the difference between a network of friends, and potential isolation.