Similarity for news recommender systems

Nava Tintarev and Judith Masthoff

University of Aberdeen, Aberdeen, UK
ntintare{jmasthof}@csd.abdn.ac.uk

Abstract. The accuracy of content-based recommender systems tends to depend on the way similarity is defined. In this paper, we will explore different ways to measure similarity for a news recommender system based on news headlines. We will compare human judgements of similarity with Lin's taxonomy-based measure and the WASP measure that uses annotated corpus data. The main aim of this work is to better understand similarity, so that it can be used to explain recommendations to users.

1 Introduction

Recommender systems using content-based filtering suggest items according to a similarity measure between items already known to be preferred by users, and new prospective items (e.g Pazzani & Billsus, 1997). This similarity measure needs to be chosen with care (Ziegler et al., 2005). An overly narrow definition of similarity between items can lead to a portfolio effect, where identical or near identical items are recommended. Examples of this are the recommendations given on the Amazon website (www.amazon.co.uk). A user looking at *Hard Times* by Charles Dickens will, for the most part, be recommended more books by Charles Dickens. On the other hand, a very broad definition of similarity will lead to poor accuracy at best, and an overload of information in the worst case.

It is important for a recommender system to be able to explain its recommendations to users, so to be transparent and scrutable (Kay, 1999; Barneveld & Setten, 2004). A recommendation can be explained in terms of the ways in which the item is similar to items the user has rated before. So, rather than a similarity measure behaving like a black box, we would like it to provide a user understandable explanation (e.g. a cricket item is similar to a football item because they are both about sports).

In this paper, we will explore the topic of similarity in the context of a news recommender system. Section 2 discusses various ways similarity is currently being measured between concepts, and extends this to similarity measures between sets of concepts. Section 3 investigates how humans judge similarity for news headlines. Section 4 investigates how well two of the similarity measures perform on the same headlines, i.e. how well they correlate with the human judgments. Section 5 presents some conclusions.

2 Similarity

2.1 Similarity between concepts

Fernández et al's (2006) recommender system bases inter-concept similarity on the distance of the concepts to their least common subsumer in an ontology. This kind of similarity measure is related to the path distance measures that have been used by a number of researchers including Lee et al (1993). This definition can be used to explain recommendations to users.

However, Resnik (1995) pointed out that the links in a taxonomy do not always represent uniform distances. He proposed a similarity measure based on the *information content* of the common ancestor(s): for concepts c_1, c_2

$$\text{Similarity}(c1, c2) = \max_{c \text{ common ancestor of } c1 \text{ and } c2} -\log \text{P}(c)$$

where P(c) = probability of c or its descendants occurring based on a corpus.

Basing his work on Information Theory, Lin (1998) also advocates the use of information content for defining similarity. He proposes for concepts c_1, c_2

$$\text{Similarity}(c1, c2) = \frac{2 \times \log \text{P}(c)}{\log \text{P}(c1) + \log \text{P}(c2)}$$

where c is the common ancestor of c_1 and c_2, and P() is defined as above.

The main difference with Resnik's definition is that Lin includes how concepts differ, not just what they have in common. Lin's definition resolves an important problem of Resnik's. Using Resnik's definition, similarity between a concept and itself varies between concepts (Richardson et al., 1994). Lin's definition results in self-similarity being 1 for all concepts.

Resnik, Lin and others have used the WordNet taxonomy to determine common ancestors. Corpus data, such as the BNC (British National Corpus) has been used to determine the probabilities. WordNet has its limitations. For instance, we considered the similarity between "invention", "painting" and "play" (in the sense of theatrical play). We found that "invention" was a lot more similar to "painting" than to "play". While "invention" and "painting" both descended from "artifact", "play" did not. It was in a completely different hierarchy, descending ultimately from "communication". Another problem with basing similarity on common ancestors is that words that seem highly related, like "Painter" and "Painting" may be very far apart in an "is-a-kind-of" hierarchy.

Instead of using a taxonomy as a basis for calculating similarity, it is also possible to use only annotated corpus data. The frequency with which concepts are used in the same grammatical structures (namely "dependency triples" that consist of a word, a grammatical relation, and a word) can be used to define similarity (e.g., Lin, 1998). Kilgarriff and Tugwell's (2001) WASPS system is based on an adaptation of this idea. We again compared the similarity between "invention", "painting" and "play", and WASPS found that painting and play were more similar to each other than either

to invention[1]. WASPS also has its limitations. It does not distinguish between word senses (in contrast to WordNet). For instance, "Royalty" seems to be far more frequently used in the meaning of "Bonus", "Premium" then in the intended sense of "Monarchs". This is likely to impact its similarity ratings.

Latent semantic analysis (LSA) (Landauer et al., 1998) takes the idea of using a corpus without a taxonomy even further: it uses pure corpus data, without even grammatical annotations. It learns from a corpus how likely words are to occur in the same contexts (sentences, paragraphs, texts). It does not just look at the frequency of word co-occurrence, but learns more hidden relationships. For instance, if words w1, w2, and w3 tend to occur together, and words w2, w3, and w4 tend to occur together, then it may induce a relationship between w1 and w4 even if they were never to occur in the same text (perhaps because they are synonyms). Because LSA does not use a taxonomy nor grammatical information, it has no difficulty learning that "Painter" and "Painting" are similar.

2.2 Similarity between sets of concepts

News headlines contain multiple words, so we need to decide how to calculate the similarity between headlines like "Painters and Inventors" and "Kings and Queens". It seems sensible to calculate Similarity(co_1 and co_2, co_3), for concepts co_1, co_2, co_3, as a function of Similarity(co_1, co_3) and Similarity(co_2, co_3). Obvious candidates for this function are the maximum, average, or minimum.

Consider headlines "Lions and Elephants" and "House cats and Airplanes". The similarity between "House cats" and "Lions" is probably quite high, as both are cats. The similarity between "House cats" and "Elephants", "Lions" and "Airplanes", and "Elephants" and "Airplanes" are likely to be a lot lower. Taking the maximum similarity does not seem the right decision, as it would make these two headlines as similar as "House cats" and "Lions". The problem becomes even worse if we use a similarity measure that does not distinguish between word senses. Consider headlines "Royalty and Palaces", and "Premiums". The similarity between "Royalty" and "Premiums" might be very high, given that "Royalty" has a word sense "Bonus". However, we believe that the similarity between "Royalty and Palaces" and "Premiums" should be very low. Based on these examples, it is tempting to use the minimum.

However, this definition would result in the Similarity between "Painters and Inventors" and "Painters and Inventors" being lower than expected, as the minimum would dictate this to be Similarity(Painter, Inventor). Using the minimum seems therefore not the right decision. Using the average would have the same undesirable effect. This leads us to using the maximum, despite our earlier reservations. However, we do not simply take the maximum. This would have resulted in headlines "Lions, Guns, Swords and Spears" and "Tigers, Chocolates, Raisins and Biscuits" to be as similar as "Lions" and "Tigers". Instead, we will consider each concept in s_1, and decide how similar it is to s_2.

---

[1] Play was 47 in ranking for Painting, Painting 108 in ranking for Play. Painting was 471 in ranking for Invention, Invention 669 in ranking for Painting.

To preserve commutivity, i.e. Similarty(s_1,s_2)=Similarity(s_2,s_1) , we will also consider each concept in s_2, and decide how similar it is to s_1. We will average over all concepts. Hence, we define

$$\text{Similarity}(s1, s2) = \frac{\sum_{co1 \in \text{Concepts}(s1)} \text{Similarity}(co1, s2) + \sum_{co2 \in \text{Concepts}(s2)} \text{Similarity}(co2, s1)}{|\text{Concepts}(s1)| + |\text{Concepts}(s2)|}$$

with $\text{Similarity}(co, s) = \max_{co2 \in \text{Concepts}(s)} \text{Similarity}(co, co2)$, for concept co and headline s

3 Experiment: Human judgement of similarity

As a basis for comparison, we first investigated how humans judge similarity between news articles, based on only headline information.

3.1 Method

Twenty-six computing students of the University of Aberdeen participated in the experiment, which was conducted online in a classroom setting. Each subject was given nine comparisons of news articles A and B. Only the headlines of the news articles were given. For each comparison, subjects were asked three questions about the similarity of the news articles using a seven point Likert scale:

1. How related is the topic of news article A to that of news article B? (scale from Dissimilar to Similar)
2. You know somebody is interested in news article A. How sure does this make you that they will also be interested in news article B? (scale from Unsure to Sure)
3. If you have read news article A, how much new information might news article B give? (scale from None to A lot)

All headlines were taken from Google news (*news.google.com*) on 13th March 2006. They originated from several English speaking news editions (e.g. in South Africa, UK, USA). We purposely did not restrict ourselves to say British articles, in order not to benefit a particular similarity measure (e.g. WASP uses the British National Corpus, while WordNet is mainly built by Americans).

Headlines were selected from various categories, such as Entertainment, Science and Technology, Sports and World news. For each headline, we also selected another two "identical" headlines, namely the two headlines given in Google news about the same topic. For instance, for headline "Finnish road transport strike ends", we also selected "Transport strike continues, mediation effort rejected Sunday night" and "Initial impact of transport strike on Finnish industry is small". To give another example, for headline "Sean Connery fit after tumour operation", we also selected "Sean Connery undergoes surgery for kidney tumor" and "James Bond on the mend". This resulted in sixteen groups of three "identical" headlines.

Subjects were given three different types of comparison:

"Identical" (I). The two headlines to compare were taken from within a group.

"Related" (R). The two headlines to compare were taken from two different groups, but within the same category, for instance, both from Entertainment.

"Different" (D). The two headlines to compare were taken from an altogether different category, for instance, one from Entertainment and one from Science.

Example comparisons of these three types are given in Table 1.

Table 1 Example comparisons of type I, S, and D.

	Headline A	**Headline B**
I	Sean Connery fit after tumour operation.	James Bond on the mend
R	Sean Connery fit after tumour operation.	Crowe's son names his brother
D	Sean Connery fit after tumour operation.	NASA relieved as probe makes orbit

To ensure that the topics chosen would not have too much impact, the sixteen groups of headlines were divided into four sets of four groups (with each four groups consisting of two pairs of groups from the same category), and subjects were randomly assigned to one of them. Subjects were presented with an I, R and D comparison for three of the groups within their set (group 4 being used to provide an R comparison for group 3). Which other headline was used in I comparisons was randomized. The order of the headlines was also randomized.

3.2 Results and Discussion

Table 2 shows the results for the three questions per comparison type. A general linear model with repeated measures showed a significant effect of comparison type for Questions 1 and 2 ($p<.001$). Pair wise comparisons showed a significant difference between the comparison types I and R ($p<.001$), as well as between R and D ($p<.01$), for both Questions 1 and 2.

Table 2 Results for Questions 1, 2 and 3 for the different comparison types.

	I comparisons		R comparisons		D comparisons	
	Mean	StDev	Mean	StDev	Mean	StDev
Question 1	2.12	1.42	4.15	1.99	6.38	1.30
Question 2	2.50	1.36	4.35	2.08	6.15	1.05
Question 3	3.69	1.59	3.54	2.12	3.77	2.78

However, no effect of comparison type was found for Question 3. Subjects' comments revealed confusion with this question. Some subjects believed "new information" in this question to mean new information overall, while others interpreted it as new information pertaining to the subjected initiated in Headline A. This may explain the lack of difference.

The significant results for Questions 1 and 2 suggest that subjects indeed thought that "identical" headlines are more similar to each other than to "related" headlines, and more similar to "related" headlines than to "different" headlines.

4 Automatic judgement of similarity

The next step was to give automated similarity measures the same task as our human subjects. We pre-processed the news headlines: we manually went through them and selected the nouns. Some nouns varied in spelling across headlines, for instance "tumour" and "tumor". Those were kept as they were. We have used "hot potato" and "car bomb" as nouns, as they are common concepts and present as such in WordNet.

4.1 Lin measure applied to WordNet

We have implemented a similarity measure between sets of concepts using the PERL interface to WordNet and implementation of Lin's measure of similarity provided by (Pedersen et al., 2004), using the BNC corpus for information content. To use Lin's measure, we needed to know which word sense to use for each of the nouns. We have manually annotated the nouns with their correct word sense in WordNet. The following concepts were not in WordNet: Russel Crowe, Crowe, Milosevic, Sean Connery, Mohali. Similarity between those concepts and other concepts would therefore be zero. We replaced "test" with "test match", as the right word sense of "test" (meaning cricket match) was not in WordNet. We used the only sense of "explorer" in WordNet ("someone who travels into little known regions"), which is not exactly right given that the "explorer" in the headline is a vehicle, not a person. We used capitals in the same way as they are used in WordNet.

The same comparisons as done by human subjects were done using Lin's similarity measure. Examples of the results are shown in Table 3, with descriptive statistics in Table 4. A general linear model with repeated measures showed a significant effect of comparison type ($p<.001$). Pair wise comparisons showed a significant difference between the comparison types I and R ($p<.001$), and R and D ($p<.05$). The Lin similarity measure was significantly (Pearson's correlation, $p<.001$) correlated with subjects' answers to Questions 1 ($r=-.56$) and 2 ($r=-.50$).

So, overall the Lin measure performs quite well. There some cases where it performs badly though. One of those is the second example in Table 3, where it gives a very low similarity to "identical" headlines. This is caused by "death" having nothing in common with "autopsy" in WordNet (they do not even have a common subsumer, as they are part of different WordNet hierarchies, death being an event, and autopsy an act). Human subjects had no problem making the link between autopsy and death as clear from their comments (e.g. "A and B share the theme of death, and are hence at least a little related"). Similarly, "stork" and "baby" are only related in WordNet as both are living beings. WordNet also does not know that "Sean Connery" has anything to do with "James Bond" ("Sean Connery is undefined, but even if he were, they would still have noting in common as "James Bond" is classified as ultimately a

"psychological feature"). A number of concepts being undefined also affects the similarity particularly of the "identical" headlines.

Table 3 Results for some example comparisons. Numbers in brackets give senses other than 1.

Type	Comparison	Lin	WASP
I	transport, strike, mediation, effort, Sunday, night(2) road, transport, strike	0.52	.50
I	reaction, death, Yugoslavia expert, autopsy	0.03	0
R	road, transport, strike doomsday(2), export, commuter(2)	0.26	.05
R	bird, flu flu, vaccine, kid	0.55	.46
D	road, transport, strike police, security(4), government, house(2), turmoil(3), Thailand	0.10	.12
D	impact(3), transport, industry police, security(4), government, house(2), turmoil(3), Thailand	0.35	.16

Table 4 Means and standard deviations using same comparisons as done by subjects.

	I comparisons		R comparisons		D comparisons	
	Mean	StDev	Mean	StDev	Mean	StDev
Lin	.47	.21	.24	.17	.13	.10
Wasp	.38	.17	.16	.18	.06	.06

4.2 WASP measure

We have implemented a similarity measure between sets of concepts in PERL using Albert Gatt's PERL interface to WASP. We have only used nouns, as we did for WordNet, although WASP could cope with adjectives. The same comparisons as done by human subjects were done using WASP's similarity measure. Examples of the results are shown in Table 3, with descriptive statistics in Table 4. A general linear model with repeated measures showed a significant effect of comparison type ($p<.001$). Pair wise comparisons showed a significant difference between the comparison types I and R ($p<.001$), and R and D ($p<.05$). The WASP similarity measure was significantly (Pearson's correlation, $p<.001$) correlated with subjects' answers to Questions 1 ($r=-.57$) and 2 ($r=-.47$). It was also significantly and very strongly correlated with the Lin measure (Pearson's correlation, $p<.001$, $r=.91$.

The performance of the WASP measure was very similar to that of the Lin measure. For instance, WASP did not find any similarity between "autopsy" and "death" either. It should be noted that we have used a version of WASP that only provides similarity between concepts and the 500 most related concepts. WASP may well perform better if no cut-off were used. We should also try whether incorporating adjectives would increase performance. However, the high correlation between WASP and

Lin is encouraging, in that WASP needs less information than Lin as we do not need to give it word senses. This would make it a lot more viable to use in a real world setting.

5 Conclusions

Both the Lin and WASP measures seem feasible options for calculating similarity in the context of real-world news headlines. WASP has the advantage of not needing word senses. However, the Lin measure is better suited for construction explanations for users. We propose a hybrid approach, in which we calculate both the WASP and the Lin measures, using word sense 1 for all words (so, that we do not have to manually annotate). We use the WASP measure as our similarity measure, but whenever the measures give similar results, use the common subsumer as used by Lin to explain the similarity to the user.

References

Barneveld, J.V. and Setten, M.V. (2004). In L. Ardissono, A. Kobsa and M. Maybury (Eds.) *Personalized digital television,* Kluwer Academic Publishers*,* pp. 259-285

Fernández, Y. B., Arias, J. J. P., Solla, A. G., Cabrer, M. R., and Nores, M. L. (2006). Bringing together content-based methods, Collaborative filtering and semantic inference to improve personalized TV. *4th European Conference on Interactive Television (EuroITV 2006)*

Kay, J, (1999) *A scrutable user modeling shell for user-adapted interaction*, PhD Thesis, Basser Department of Computer Science, University of Sydney, Australia.

Kilgarriff, A., and Tugwell, D. (2001). Word sketch: Extraction and display of significant collocations for lexicography. In *Proceedings of the Collocations Workshop* in association with ACL-2001, Toulouse, France.

Landauer, T.K., Foltz, P.W., and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*, 259-284.

Lee, J. H., Myoung H. K., and Yoon J. L. (1993). Information retrieval based on conceptual distance in IS-A hierarchies. *Journal of Documentation, 49*, 188-207.

Lin, D. (1998). An information-theoretic definition of similarity. *Proceedings of the Fifteenth International Conference on Machine Learning*, 296 – 304.

Pazzani, M. and Billsus, D. (1997). Learning and revising user profiles: The identification of interesting websites. *Machine Learning, 27*, 313-331.

Pedersen, T., Patwardhan, S, and Michelizzi, J. WordNet::Similarity – Measuring the relatedness of concepts. *Proceedings of the 19th national conference on Artificial Intelligence (AAAI-04)*. San Jose, CA., pp. 1024–1025

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pp448-453, Montreal.

Richardson, R., Smeaton, A.F., and Murphy, J. (1994). Using WordNet as a knowledge base for measuring semantic similarity between words. Working paper CA-1294, Dublin City University, Dublin, Ireland. http://www.computing.dcu.ie/research/papers/1994/1294.pdf

Ziegler, C., McNee, S.M., Konstan, J.A. and Lausen, G. (2005). Improving recommendation lists through topic diversification *Fourteenth International World Wide Web Conference.*