Information Research, Vol. 5 No. 1, October 1999

Using newsgroup headers to predict document relevance

Christopher Brown-Syed and William Morrissey
Library and Information Science Program
Wayne State University
Detroit, Michigan, USA

ABSTRACT

Describes a pilot study of Usenet Newsgroup postings about Philosophy. The aim of the research is to arrive at a set of predictors of the quantity and usefulness of documents posted on the Net with the aim of assisting educators in the development of curricula, and LIS practitioners in the assessment of networked resources. In particular, we wished to see whether relationships existed between the layout of digital artefacts and the usefulness and reliability of their content.
The current study suggests that a strong correlation exists between the email header elements stating the number of cross postings to multiple newsgroups and the relevance of a posted document to its purported topic. Articles deemed "useful", namely those which exhibited facility with technical terminology and methods appropriate to the field of discussion, also appear to be briefer or less verbose. We suggest that the number of cross-postings, and the number of lines in a message, are better predictors of the usefulness of message content than are other header elements such as the purported subject, or the sender's Internet domain of origin.

Introduction and Statement of problem

The process of scholarly publishing and editing, including the peer review process, has been in existence for many generations. Publishers have adopted certain conventions, such as title pages, contents pages, descriptions of authors' credentials, references, and indexes, which assist librarians and other researchers who wish to assess their reliability. The layout of books allows immediate categorization - as "mass-market", "trade", or "scholarly" material.

Scholars are thus able to appraise monographs and serials quickly, based on their physical appearance or packaging, before actually examining their texts. In some cases, the reputation of the book publisher, or the journal's editorial board, may be recommendation enough. While no scholar can claim to provide the absolute 'truth', the editorial process ensures that the contents of printed materials have been given reasonable scrutiny prior to their release.

The same is true of audio visual media, such as sound recordings, films, and videotapes, which have equivalent characteristics. For instance, a documentary film can usually be distinguished from a dramatic work, merely by examining its citation or by looking at its credits.

However, digital media present evaluation problems which are to some degree inherent in their design. There is no "cover" nor "title page" to an email message, and rarely are credits and references given for Web pages. Digital documents are volatile, transient, and multitudinous.

Accordingly, in these pilot studies, we sought to examine the headers of email type postings to Usenet Newsgroups on academic topics, and the amounts of personal self-revealing information provided by selected groups of academic Web authors. The primary group of people studied were those purporting to hold credentials in, or to be seriously interested in the topic of Philosophy. We wished to determine, in the first instance, whether there were any email header elements which might reliably forecast the appropriateness of the texts, and in the second, to determine how much we could determine about authors' bona fides by examining their Web pages.

Review of the Literature

Within the literature of librarianship, many of the articles on internet use have been what might be termed "use studies". These include a series of publications emanating from Australia. They include works by Applebee et al. (1997), Bruce (1995 and 1994), Pascoe et al. (1996), and Organ and McGurk (1996). Blinko (1996) conducted a study of staff and students at a British university, while McClure (1994) discussed the broader impacts of the Net on academia in the United States. Other articles focus upon the searching behavior of users, and upon reference services and use of the net by practitioners. Clayton et al. (1996), discuss the problems associated with user surveys conducted over the Net. Having conducted an email survey ourselves (Brown-Syed and Witzke, 1996), and having experienced quite a low rate of return, we did not attempt to use similar methods in this phase of our research. It is partially because email surveys are not terribly representative that we wished to study the documents themselves. (Brown-Syed and Witzke, 1996).

Studies of related Internet services, especially of the World Wide Web, are more prevalent. Our 1996 study of traffic on a Bitnet Listserv designed to provide peer support for users of Unesco's Micro CDS/ISIS library automation software (Brown-Syed & Witzke, 1996), sought to determine the usefulness of the information contained on the list, and the reasons for its use. This study included a user survey and analysis of the content of postings. It concluded that, while the list was indeed fulfilling its function of user support, a striking number of postings were of limited use, or offered seemingly conflicting advice. If this was the case in a moderated mailing list, how would posting patterns in unmoderated Usenet Newsgroups compare?

Like Bitnet lists, (Listservs), Usenet News predates the global Internet. In the 1980s, it was employed heavily by commercial, educational, and military users, and offered non-profit organizations an inexpensive vehicle for the exchange of ideas and information. Like Listservs, Usenet Newsgroups can be moderated by human or mechanical editors, but by far the majority of them are unmoderated. Because of the ease with which users can subscribe and unsubscribe to Newsgroups, they are perhaps more vulnerable to spurious or misdirected messages.

In the 1980s, Newsgroups were used for scholarly and technical communication, as well as for recreation. Netiquette guides like Brad Templeton's "Emily Postnews" documents (which appears from time to time in *news.announce.newusers*), informed users that there was a place on the Net for any sort of conversation, but that it was up to users to ensure that they discussed designated topics in appropriate groups. Ed Krol's classic *Whole Internet Catalog* explains the hierarchy of Newsgroups, and the complicated voting procedure used to create new ones (Krol, 1995).

In 1998, T. Matthew Ciolek, of the Research School of Pacific and Asian Studies, Australian National University, conducted an "impromptu" survey of some 1700 academics on the Net. This survey was aimed at ascertaining the time spent in various online activities. Ciolek observed: "Activities on the Net are twofold: (1) traversing the system and making use of its various information resources; (2) active research and construction work aimed at provision of networked information resources. Not surprisingly, construction work was less common (and less time-intensive) of the two." (Ciolek,1998). The results of this survey, while interesting, do not assist us greatly in determining just who is constructing which resources, and how useful, reliable, or effective those resources might be. However, they do suggest that network creation activities increase with user experience: Ciolek's observations on construction of personal pages are congruent with the results of the Oct-Nov 1997 Survey (GVU 1997) where 10,108 participants reported that 46% of them have created a web page and where was found that "the percentage of respondents creating web pages increases with the length of their online experience." (Ciolek, 1988). If one wished to apply this logic to Usenet News postings, an author's frequency of posting and cumulative number of posts might be considered analogous. Tools like *DejaNews* have allowed reaearchers to perform citation searches or topical searches, but because of the volume of traffic, most sites do not keep copies of the News for lenghty periods.

When we began our investigations, personal Web pages on the Net had been the subject of only one study of which

we were aware. Bates and Lu (1997), who reported no similar studies, performed a content analysis of 114 personal Web pages derived from an online directory available through Netscape's menus. The purpose of their study was not to determine grounds for accepting or rejecting information, but rather, to categorize the types of pages found with respect to metaphors from other media. Its goal was to determine whether certain types of information were more apt to be present consistently, and to see whether any conclusions about stylistic elements and the cultural function of personal Web pages emerged. Quoting December & Randall (1994), they defined "home" pages as designated entry points to local Web sites. Operationally, in this study, a home page was defined as the first screen of information that appears upon entering the URL address (drawn from the People Page directory) of a Web site.... (Bates & Lu, 1997). However, as analysis of the authors' own web site access logs demonstrate, the nominal "front" or "home" pages are quite apt to be bypassed by users who use search engines or direct links to locate specific documents on the target sites (Brown-Syed, 1999a).

Table 1. demonstrates how access to pages not linked from an academic site's home page can constitute the majority of traffic from bona fide off-campus visitors during a two fairly typical periods toward the middle of an academic term. The high number of hits on the "King Arthur" material can perhaps be accounted for simply. Of several thousand pages devoted to this supposed monarch, the page ranks highly on search sets retrieved from AltaVista, and other engines. As well, other Arthurian sites link to this short bibliography *cum* Weblog. Even during an exceptionally busy time for students, hits to the home page constitute only about a tenth of all visits to the site.

29 Sept - 05 Oct 1999. (N = 1539).

King Arthur	**117**	**11.5%**
Home Page	**98**	**6.4%**

21-27 Sept 1999. (N = 2716).

King Arthur	268	9.9%
Home Page	260	9.6%

Table 1. External visits to valinor.purdy.wayne.edu

The assumption that a site's popularity is an indicator of its quality or relevance is problematic. Moreover, results can be easily skewed. The publication of the current article may in itself skew future results, as other researchers follow the links it contains.

Since 1994, the Graphic, Visualization, & Usability Center (GVU) located at Georgia Tech, has conducted surveys of the Net aimed at collecting demographic data. (GVU, 1999). Descriptions of several papers derived from these surveys are available online, however, the documents are "preprints" and may not be cited nor quoted. The URL for the GVU study appears in the attached References. The GVU surveys do not focus upon the academic use of the Net. However, some of their findings are of interest, for instance, while the October 1997 GVU survey found a majority of Net users were over 50 years old, for the first time, among the cohort of newer users, there was a higher ratio of females to males.

A demographic study based upon responses of about 1800 network users, had been performed by the Canadian government in 1996. By comparison, it had determined: "[T]he majority of respondents (72.6%) were between the ages of 25 and 54 years of age.... The majority of all respondents were males (76.6%), 23.4% were females." (Canada, 1996, summarized in Canada, 1999b). This survey determined that most network users were highly educated, middle class, and living in heavily populated and economically prominent areas of the country. The survey was not limited to Web use. It contained questions about other services, such as telnet, ftp, and discussion groups. (Canada, 1999a). However, it did not concentrate upon the creation of networked documents, nor upon the scholarly use of the Net, but upon users of government Web sites. Moreover, it depended upon an admittedly skewed sample - only people who found the survey could respond to it. Surveys like these give us insight into the characteristics of Net users, and may provide benchmarks against which to compare the characteristics of the cyber-scholar population.

With the advent of the global Internet and the entry of commercial network service providers, and the popularity of

the World Wide Web, older services such as mailing lists and Newsgroups have been largely ignored, perhaps deservedly so. Some of the older Newsgroups have lost their hitherto professional character, and appear to have become the haunts of users who have never read about Newsgroup hierarchies and Netiquette. Some groups are now completely given over to mass-mailed advertisements, despite the former strict adherence to rules which allowed advertisements only in certain groups. Indeed, as this article went to press, the home page at "DejaNews", now re-launched as deja.com had all but completely de-emphasized the organization's original role as a Usenet news search engine facility, and had been largely given over to online shopping.

These developments do not necessarily spell the end of Usenet. Rather, they may herald a return to the facility's halcyon days as a vehicle for serious scholarly and scientific exchange. The NASA sponsored news group, sci.space.news, for instance, carries up to date press releases and mission itineraries, and contains almost no spurious messages.. The group is moderated, and reader comments are re-directed to another group, called sci.space.policy.

Methods

We felt that the usefulness of message contents could be functions of the posters' proper use of jargon appropriate to a given field, their provision of references (whether direct or indirect), and their stylistic familiarity with the medium. We subsequently adapted the methods used in this study in our examination of Web authorship. After examining several groups on the screen, we were able to arrive at several specific hypotheses.

In particular, we felt that the use of technical jargon appropriate to a field, and the contribution of authoritative or verifiable material in postings about that field, might be predicted from the path, from, subject, and physical extent fields of the posting headers. The manner in which we operationalized the dependent and independent variables is outlined below.

Our dependent variable, "Usefulness", was determined by collapsing measures of the use of appropriate technical jargon, and of evident relevance to the expressed title or topic given in the "subject" line.

Independent variables proposed included: the number of cross-postings to other groups, the length of the documents themselves, the domains of origin of the senders, and their use of actual rather than proxy addresses.

Two Usenet Newsgroups related to philosophy were chosen for this study. We chose these groups because we were familiar with technical or professional philosophy , and felt we would be able to identify direct or oblique references to its recognized practitioners and arguments. We chose philosophy groups because we felt, due to our previous academic degrees, that we would be competent to recognize the terms and concerns of professional philosophers when they occurred in the texts of News documents.

Beginning with the first article carried on the University's news server, we extracted every tenth article from two related Newsgroups, to a total of twenty articles per group. By examining the documents, we extracted the following information:

1. domain of origin (com, mil, edu, etc.)
2. "real life" name and email address of poster
3. actual machine of origin (determined from the Path line)
4. user-declared site of origin (the From line)
5. number of lines of text in the message
6. number of newsgroups to which the message was cross-posted

As well, we examined attached ".signature" files, when these were provided. Some experienced users use the News reader's capabilities to attach contact information automatically to outgoing mail. Typical signatures include postal addresses, phone, alternate email, or fax numbers, along with affinity statements, such as "Associate Professor, Such and Such University". However, not all users are aware of the signature mechanism, and some experienced users append very brief ones, often humorous rather than informative in nature.

Interestingly enough, during our preliminary search for groups to study, we discovered a fraudulent professor. The poster, evidently a computer programmer in Melbourne, Australia, passes himself off as a Professor of Developmental Psychology at a non-existent university, evidently for the purpose of eliciting information about

sexual behaviour from unsuspecting teen agers. Accordingly, we felt that a more detailed examination of signature files would best be left to a follow-up study.

To operationalize the dependent variables, we applied two measures to the document contents. A technical jargon variable and a relevance variable were constructed.

Each document was assigned a technical language value from 0 to 2. If a poster clearly referred to commonly known philosophical arguments, named specific philosophers, or constructed formal philosophical arguments, the document was assigned a value of 2. If the poster appeared to be concerned with a traditional problem or branch of philosophy, but did not employ professional methods or references to technical philosophy, the document was assigned a value of 1. If the user's posting had no philosophical content whatsoever, it was given a zero.

Similarly, we compared document contents to the stated subject of the posting as provided in the header. A document which clearly advanced the topic of discussion was assigned a value of 2, while one only moderately on topic was given a 1, and a value of zero was assigned to documents which had no bearing whatsoever on the topic at hand.

Realizing that staying on topic and employing appropriate technical language or methods could be interpreted as two aspects of a document's relevance, we later combined the two original independent variables, labelling the composite variable "usefulness". We then used statistical tests to determine the likely influence of the individual independent variables, derived from the document headers, as well as combinations of them, upon each dependent variable in turn, and upon the composite variable "posting relevance". The data were analysed using SPSS 9.0, with the results shown in Table 2:

		lines	**techno**	**subj= disc**	**x-posts**	**useful-ness**
lines	Pearson correlation	1.000	-.279	-.107	.738**	-.222
	Sig. 2-tailed	-	.094	.528	.000	.187
	N	38	37	37	38	37
techno	Pearson correlation	-.279	1.000	.678**	-.401*	.935*
	Sig. 2-tailed	.094	-	.000	.014	.000
	N	37	37	37	37	37
subj=disc	Pearson correlation	-.107	.678**	1.000	-.209	.894*
	Sig. 2-tailed	.528	.000	-	.215	.000
	N	37	37	37	37	37
x-posts	Pearson correlation	.738**	-.401*	-.209	1.000	.345*
	Sig. 2-tailed	.000	.014	.215	-	.037
	N	38	37	37	38	37
usefulness	Pearson correlation	-.222	.935**	.894**	-.345	1.000
	Sig. 2-tailed	.187	.000	.000	.037	-
	N	37	37	37	37	37

Table 2: Correlations between variables

In Table 2, drawn from a sample of postings in philosophy newsgroups, we see how a document's scores on use of appropriate technical language, the tendency of its contents to reflect the stated subject of discussion, and therefore its "usefulness", are inversely proportionate to the number of lines contained therein, and to the number of groups to which the document has been cross posted. While we observe strong relationships between pairs of variables, such as "use of technical language" and the tendency to "stay on topic", among the strongest relationships, and therefore, among the most reliable quick indicators of a document's likely relevance, is the number of groups to which the message has been cross posted. This fact is particularly useful in quick reference settings, because it is easily determined from the document header, and requires no facility with technical terms particular to a discipline. Similarly, the number of lines in a posting appears to correlate well with the number of cross posts - the most verbose postings tend to be the ones most widely distributed.

Discussion

As the following summary statistics and Table 2 demonstrate, our investigation of email header elements suggests that a cursory examination of the header elements "lines" and "x-posts" can indeed be used is a rough predictor of a document's relevance to the announced topic of a Newsgroup. However, further research is required due to some ambiguities in the current research design. These will be explained below.

Ambiguity arises first and foremost in the determination of relevance, which we have expressed in terms of a collapsed variable called "usefulness". Our samples were drawn from two specific Newsgroups, and we have considered all articles with respect to their relevance to those two points of entry. However, any one user may have post a message to another group, and for various reasons, chosen to copy the message to one of the groups we sampled. Further studies might attempt to analyse the priorities given to groups named in the "x-posts" header element. Such research would have to cope with the fact that "replies" to postings are often made indiscriminately. A user sees a message, touches the "follow-up" key, and does not necessarily check to see how many groups the reply will affect.

This "quick trigger finger" behaviour can be embarrassing, and "noisy", especially if the original poster was merely seeking answers to an open question, and had not really determined the "most proper" group in which to post. It can lead to a sort of institutionalized irrelevance in the News. Once a thread of discussion begins, subsequent contributors may perpetuate an initial misdirection unintentionally.

Another ambiguity may obtain from the broad focus of the groups selected, and from the fact that words like "metaphysics" have popular as well as technical meanings. The same may not be true of other topics. For instance, the term "personal water craft" has a precise meaning, and it is unlikely that users would post messages in a group devoted to this topic unless they were aware of that meaning.

In future studies, we hope to draw representative samples from other newsgroups whose known or stated functions are professional or academic, and also from those whose users seek primarily to exchange recreational or avocation information. For example, genealogists have a tightly controlled, terse, and concise way of requesting information from other genealogists. Almost all of the postings in genealogy Newsgroups exhibit voluntary conformity to standards which have evolved over the past decade or so. It is possible that amateurs of other fields, such as computing, photography, skiing, or other hobbies, are more productive in their use of the medium than are users interested in more intellectual pursuits like philosophy. Because genealogy, for instance, has its own vocabulary, which tends to dissuade outsiders, chance users may feel less inclined to post rambling or badly-formed requests to genealogy lists than do those whose interest in philosophy is avocation rather than professional. The same may be true of other scholarly mailing lists and Newsgroups, which, by their very natures, invite speculation from people in other fields of endeavour.

A much larger sample, using data drawn from a variety of discussions on various fields and reflecting various disciplines, would be required before any sweeping conclusions could be drawn. However, with the volume of traffic in Newsgroups, and the number of Newsgroups both increasing exponentially, any sampling method which was based on an educated guess about the population would likely prove fruitless. However, our pilot study suggests that of all the elements in standard Newsgroup email headers, number of cross postings and length of the submission in lines are generally good at predicting the likely usefulness of postings. If librarians or researchers are in a hurry, these elements may help them identify relevant documents quickly.

It appears that we were correct in assuming that some indication of the value of a newsgroup posting can be ascertained from the document headers. While we recognize the limitations of this pilot study, we feel that it is possible to construct measures of document relevance and to predict their occurrence based on header information.

We attach one important caveat to this suggestion: in order to construct such a measure, one must be somewhat familiar with a given field. This study made no at tempt to confine itself to "established" newsgroups - ones which had passed the complicated voting procedures outlined in Krol (1995) and conducted in the Newsgroup news.announce.newgroups. Groups established outside the formal process, "alternate" groups, often have no charters or rules of appropriate conduct. Future studies should compare the traffic in "alternate" groups, with "official" groups established by calls for votes. As well, the traffic of "talk" and "miscellaneous" groups should be compared with that in groups with more restrictive applicability, such as those in the "science", "computing", or "society" hierarchies.

As well, there is some surface evidence from the sample that a user's frequency of posting, and the phenomenon of discussion "threads" or topics of conversation, may have considerable bearing upon the relevance of postings.

Finally, it would appear that our measure of document relevance remains valid regardless of the actual name of the newsgroup or its presence in or absence from the hierarchy of official groups. When we examined sci.astronomy.research, for instance, our University server carried only five articles. All used the jargon of astrophysics appropriately, and all discussed matters referred to in their headers. Conversely, the group alt.sex, established when a discussion group for human sexuality was defeated in the call for votes process, was originally set up for serious discussions of sexuality. It now scores poorly, according to our measure, since it is given over almost wholly to advertisements. A more careful examination of this measure of document relevance would have to be undertaken before it could be applied to general discussion groups with "talk" or "misc" in their names, since it might prove hard to identify appropriate technical jargon in these largely amateur, rather than professionally dominated, groups.

Conclusions and Suggestions for Future Research

Our investigation of Usenet Newsgroups suggested that brevity and focus of articles could indeed be determined by examining the header elements "lines" and "x-posts". Our investigation of the personal Web pages of academics suggests that those earlier in their careers tend to reveal more about themselves, hence providing more clues to credibility, and at the same time, to produce a greater volume of digital artefacts. We feel that further research into potential means of ascertaining the reliability of documents on the Net is indeed critical for information intermediaries such as librarians, archivists, or news reporters, and that the absence of editorial control on the Net makes it critical that such "hallmarks of authenticity" be determined. Despite initiatives such as Metadata, information about Net authorship is provided voluntarily for the most part, and we feel that further studies to facilitate the rapid assessment of such information are essential.

References

- Applebee, A. et al. (1997). Australian academic use of the Internet. Internet Research v.7 no.2 ('97) p. 85-94.
- Bates M. & Lu Shaojun, (1997) An exploratory profile of personal home pages: content, design, metaphors. Online & CDROM Review v. 21 (Dec. '97) p. 331-40
- Blinko, B. (1996). "Academic staff, students and the Internet: the experience at the University of Westminster." *Electronic Library* v. 14 (Apr. '96) p. 111-16.
- Brown-Syed, C. (1998). "SOS Calls, Breaking Stories, Network Disinformation, and the Process of Scholarly Communication: Implications for Information Intermediaries." Information Science at the Dawn of the Next Millennium. Ottawa: CAIS/ACSI Conference, (Proceedings), 3-5 June, 1998.
- Brown-Syed, C. (1999)."Back Door Entries, Invisible Ink, and False Drops on the Web: an Interim Research Note" *Information Research: an electronic journal* Volume 4 No 3 February 1999. Available: http://InformationR.net/ir/4-3/paper58.html
- Brown-Syed, C.; Witzke, K. (1996). "E-mail for Development: An Exploratory Study of the CDS-ISIS Distribution List." *Interdisciplinary Conference on the Evolution of World Order: Building a Foundation for Peace in the Third Millennium*. Ryerson Polytechnical University, June 6-8, 1977. (Proceedings). Toronto: Ryerson Polytechnical University and the Caledon Centre for Culture and Education of SGI Canada, 1997.
- Bruce, H. (1995). Internet and academic teaching in Australia. Education for Information v.13 (Sept. '95) p. 177-9.
- Bruce, H. (1994). Internet services and academic work: an Australian perspective. Internet Research v.4 (Summer '94) p. 24-34.
- Canada, (1999a). "Who is Using the Internet, What for and How often"? *Government of Canada Internet Guide*. Available: http://canada.gc.ca/programs/guide/1_1_3e.html
- Canada, (1999b). *Federal Government Internet Usage Survey Results*. Online. Available: http://canada.gc.ca/programs/guide/5_5e.html
- Ciolek, M., 1998."The Scholarly Uses of the Internet: 1998 Online Survey". Available: http://www.ciolek.com/PAPERS/InternetSurv/ciolek
- Clayton, P. et al (1996) "Email surveys: old problems with a new delivery medium". *LASIE* v.27 (June '96) p. 30-9.

GVU (1999). Graphics, Visualization, and Usability Center. "GVU's Eighth WWW User Survey". Available: http://www.cc.gatech.edu/gvu/user_surveys

- Krol, E. (1994). *The whole Internet catalog user's guide and catalog*. (Second edition). Sebastopol, CA,. O'Riley & Associates Inc.
- McClure, C. (1994). So what are the impacts of networking on academic institutions? *Internet Research* v. 4 (Summer '94) p.2-6.
- Pascoe, C. et al. (1996). Tidal wave or ripple? The impact of Internet on the academic. (Bibliographic essay). *Australian Library Review* v.13 (May '96) p. 147-55.
- Organ, M. (1996). Surfing the Internet and academic research: what use for historians? *Australian Academic & Research Libraries* v. 17 (Mar. '96) p. 31-9. 14 oct 98
- *News.announce.newusers*. [Electronic discussion group].

Appendix

Measures of the relationshps among number of cross-postings, lines of messages, and document usefulness. The following results were obtained using SPSS 9.0.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
x-posts * usefulness	37	92.5%	3	7.5%	40	100.0%
lines * usefulness	37	92.5%	3	7.5%	40	100.0%

x-posts * usefulness

Directional Measures

			Value	Asymp. Std. Error(a)	Approx. T(b)	Approx. Sig.
Nominal by Nominal	**Lambda**	**Symmetric**	.159	.104	1.439	.150
		x-posts Dependent	.160	.090	1.695	.090
		usefulness Dependent	.158	.145	1.014	.311
	Goodman and Kruskal tau	**x-posts Dependent**	.138	.040		.193(c)
		usefulness Dependent	.218	.046		.142(c)
	Uncertainty Coefficient	**Symmetric**	.258	.050	4.425	.235(d)
		x-posts Dependent	.230	.047	4.425	.235(d)

		usefulness Dependent	.294	.059	4.425	.235(d)
a Not assuming the null hypothesis.						
b Using the asymptotic standard error assuming the null hypothesis.						
c Based on chi-square approximation						
d Likelihood ratio chi-square probability.						

Symmetric Measures

		Value	**Approx. Sig.**
Nominal by Nominal	**Phi**	.844	.334
	Cramer's V	.422	.334
	Contingency Coefficient	.645	.334
N of Valid Cases		37	
a Not assuming the null hypothesis.			
b Using the asymptotic standard error assuming the null hypothesis.			

lines * usefulness

Directional Measures

			Value	**Asymp. Std. Error(a)**	**Approx. T(b)**	**Approx. Sig.**
Nominal by Nominal	**Lambda**	**Symmetric**	.283	.094	2.614	.009
		lines Dependent	.088	.084	1.014	.311
		usefulness Dependent	.632	.150	2.820	.005
	Goodman and Kruskal tau	**lines Dependent**	.101	.005		.752(c)
		usefulness Dependent	.708	.019		.743(c)
	Uncertainty Coefficient	**Symmetric**	.446	.041	8.300	.997(d)
		lines Dependent	.312	.035	8.300	.997(d)
		usefulness Dependent	.779	.052	8.300	.997(d)
a Not assuming the null hypothesis.						
b Using the asymptotic standard error assuming the null hypothesis.						
c Based on chi-square approximation						
d Likelihood ratio chi-square probability.						

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	1.745	.466
	Cramer's V	.872	.466
	Contingency Coefficient	.868	.466
N of Valid Cases		37	
a Not assuming the null hypothesis.			
b Using the asymptotic standard error assuming the null hypothesis.			

How to cite this paper:

Brown-Syed, Christopher & Morrissey, William (1999) "Using newsgroup headers to predict document relevance." *Information Research*, **5**(1) Available at: http://informationr.net/ir/5-1/paper64.html

Contents **Home**