A Fuzzy Web Surfer Model

B. L. Narayan and Sankar K. Pal
Machine Intelligence Unit,
Indian Statistical Institute,
203, B. T. Road,
Calcutta - 700108, India.
E-mail:{bln_r, sankar}@isical.ac.in.

Abstract

A novel web surfer model, where the transitions between web pages are fuzzy quantities, is proposed in this article. Such a model is appropriate when the links between pages are imprecise. The theoretical aspects of modeling the uncertainty associated with links are discussed. The advantages and limitations of the proposed methodology, which is based on the theory of fuzzy Markov chains, are described. Situations where fuzzy web surfer models are appropriate compared to existing models are highlighted.

1. Introduction

The World Wide Web is a complex entity, with billions of web pages and many more links between them. That there are new pages constantly being added, some old pages being deleted, and existing pages being changed all the time, adds to the complexity of the situation.

As the web consists of pages created by millions of individuals, there is a wide variety of authoring styles. Most present day content and link analysis algorithms are robust against differences in fonts, colors, *etc.*, which are mostly ornamental. Other algorithms can even withstand, to some extent, malicious manipulation of content and links. However, they are sensitive to whether the information is contained in a single document or is spread out in a collection of documents. For the sake of uniformity during comparison, either some small documents should be combined together or some large documents be split. The problem of detecting which documents should be combined together has been dealt with in [10]. The approach of combining together documents is largely useful for retrieval purposes, whereas for content and link analysis, splitting large documents into smaller coherent pieces of text seems more appropriate [10]. So, in the present context, information present in a single web page may be artificially divided into a collection of web pages. This division introduces an uncertainty in the page boundaries as well as the targets of hyperlinks.

A variety of web surfer models exist which model the sequence of web pages a surfer follows as a Markov process. The transition probabilities are obtained by considering the number of links in each page. Here, it is assumed that there is no uncertainty in the given web pages or the transition probabilities. In practice, this is not the case. This imprecision may be modeled with the help of fuzzy sets, or in particular, by fuzzy numbers. This forms the basis for the present investigation, where we extend existing surfer models to fuzzy surfer models. Since we now deal with fuzzy numbers, Markov chain theory is replaced by fuzzy Markov chain theory, which employs the $\max$-$\min$ algebra instead of the usual addition and multiplication operations. These models may be employed, among other things, to compute ranks of web pages, which we call FuzzRanks.

Fuzzy web surfer models so defined, apart from being able to handle fuzziness in various aspects, inherit the advantages of fuzzy Markov models, namely, robustness and finite convergence. Robustness is a very important aspect because it means that small changes in the transition matrix would not change the results drastically. Its significance arises from the fact that the transition matrices are not known beforehand and are estimated during the analysis phase, and so, (slightly) different methods of estimation, may lead to immensely dissimilar results. As a consequence, FuzzRank is expected to be more stable as compared to PageRank. FuzzRank reflects the belief of a surfer being on a page, and cannot fluctuate to extreme cases as in the case of probabilistic models.

The conditions when FuzzRank would be independent of the initial state of the surfer remain elusive, just as in the case of fuzzy Markov chains. This need not necessarily be a drawback, as it might be appropriate for the FuzzRanks to be dependent on where the surfer had started surfing, as in the context of web communities.

This article is organized as follows. Section 2 discusses

the preliminaries such as web surfer models and fuzzy Markov chains. We make use of these components to describe fuzzy web surfer models in Section 3, which we begin with a motivational example, and end with a discussion of the merits and limitations of the methodology under investigation. In Section 4, we look further research in this direction.

2. Preliminaries and Background

We now provide a brief background on web surfer models and fuzzy Markov chains.

Surfer models model possible surfing patterns of users to infer about various important properties of the web, such as the ranks and categories of web pages. Several surfer models may be found in the literature [2, 9, 13, 8, 11, 7], and they differ from one another in terms of how the surfer is assumed to transit from one page to another. The sequence of web pages being visited is modeled as a Markov chain, and some properties of the web are interpreted from appropriate distributional properties of the chain. For example, in the random surfer model [2], the stationary distribution is considered to be the PageRank vector.

Fuzzy Markov chains are an alternative to the classical Markov chains, where the addition and multiplication operations are replaced by the max-min algebra. Thus, the transition law of classical Markov chains,

$$\begin{aligned} P(X_{n+1}=j) &= \sum_{i\in S} P(X_{n+1}=j|X_n=i)P(X_n=i) \\ &= \sum_{i\in S} p_{ij}\times P(X_n=i),\ \forall\, j\in S, \end{aligned} \quad (1)$$

becomes

$$\mu_{n+1}(j) = \max_{i\in S}\{\min\{q_{ij},\mu_n(i)\}\},\ \forall\, j\in S \quad (2)$$

in the fuzzy case, whereby the addition and multiplication operations of Eq. 1 are replaced by the max and min operations, respectively, in Eq. 2. Here, S is the set of states that the chain may assume (or the web pages under consideration), p_{ij} and q_{ij} denote the probability and belief value, respectively, of a transition from i to j, and $\mu_n(i)$ is the belief value of the surfer visiting page i at time n. p_{ij} and q_{ij} are assumed to be independent of n. $P=((p_{ij}))$ and $Q=((q_{ij}))$ are called the probabilistic transition matrix and the fuzzy transition matrix, respectively.

The powers of the matrix Q are defined as

$$Q^{n+1} = Q\circ Q^n, \quad (3)$$

where, $\circ$ denotes the matrix multiplication in the max-min algebra. Thus, Eqs. 1 and 2 may be rewritten in matrix form as

$$\mathbf{r}^{(n+1)} = \mathbf{r}^{(n)}P = \mathbf{r}^{(1)}P^n \quad (4)$$

and

$$\boldsymbol{\mu}^{(n+1)} = \boldsymbol{\mu}^{(n)}\circ Q = \boldsymbol{\mu}^{(1)}\circ Q^n, \quad (5)$$

with $\mathbf{r}^{(n+1)}$ and $\boldsymbol{\mu}^{(n+1)}$ denoting, respectively, the probability and belief (row) vectors of the surfer's location at time $n+1$. Unlike the case of classical Markov chains, whenever the sequence of matrices $\{Q^n\}$ converges, it does so in finitely many steps to a matrix Q^τ. If it does not converge, it oscillates with a finite period ν starting from some finite power τ [1].

When $\{Q^n\}$ converges to a non-periodic solution Q^τ, the associated fuzzy Markov chain is called aperiodic and Q^τ is called the limiting fuzzy transition matrix. A fuzzy Markov chain is called ergodic if the rows of Q^τ are identical. This definition is again similar to that of classical Markov chains, but the necessary and sufficient conditions for ergodicity are not known in the fuzzy case [1].

3. Fuzzy Web Surfer Models

3.1. Motivation

We provide a real life example to demonstrate that links between pagelets need to be considered, in which case, a fuzziness is naturally associated with the presence of a link. Fig. 1 shows a portion of the web page located

> Professor Jon Kleinberg's page on Algorithms for Information Networks at Cornell University that is the defacto authority resource on Hubs and Hub theory: http://www.cs.cornell.edu/home/kleinber/

Figure 1. A portion of a web page at Webmasterworld with a link to Jon Kleinberg's home page

at `www.webmasterworld.com/forum10003/428.htm` (cached version at `www.isical.ac.in/~bln_r/428.htm`), which contains a link to Jon Kleinberg's home page `www.cs.cornell.edu/home/kleinber/index.html` (cached version at `www.isical.ac.in/~bln_r/klein.html`). This link leading to Kleinberg's home page provides no more information than its URL. The home page under consideration has two named sections, namely *Papers* and *Links*, and there is an introduction above it. It is clear from the context that the above mentioned link indeed refers to the *Papers* portion of the page. In addition, the *Papers* section is further subdivided according to the topics of the papers, but the subsections are not named. Had they been named, we may once again conclude that the link in question actually leads to the *Web Analysis and Search: Hubs and Authorities* subsection.

We observe the following from this example:

- a link to a web page may in reality be referring to just one or more pagelets, and not the whole page itself. Resolving which pagelet is referred to by a link needs contextual information, and yet, this may not be precise.
- a page may have to be artificially divided into pagelets or sections, to avoid the weight attributed by a link to one pagelet spilling over to other pagelets. As mentioned earlier, this is required for fair comparison during retrieval because, although this particular link is for a small portion of the page, the contents of the rest of the page benefit from it, thus enjoying a better status as compared to similar content elsewhere.

It may be noted that it is not claimed that one or the other is necessarily better, because some systems may assign more weight to more content, whereas, others may penalize it. All that is being argued for is that such disparities may lead to diverse results, and need to be addressed at an early stage of link and content analysis.

We now formulate a basic methodology for fuzzy web surfer models.

3.2. Formulation

We assume that the given web pages are all split into the appropriate number of pagelets [12, 6], and each of them is referred to as a page. We label these pages from $\{1, 2, \ldots, N\}$

We propose the methodology for fuzzy web surfer models by imitating that of existing surfer models. Similar to the concept of PageRank, we define the concept of FuzzRank, where the objective is to compute, for each given web page, a value which reflects the belief that a web surfer would be on that page. This value is proportional to the belief that the surfer would be on one of its backlinks. Similarly, associated with each link in a page, there is a fuzzy number that indicates the belief that this link would be followed, given that the surfer is on that page. These constitute the fuzzy transition matrix.

Formally, we are interested in computing $\mu(i)$ for each page i, which is the unconditional belief that a surfer would be on i. Whenever the fuzzy transition matrix Q corresponds to an aperiodic fuzzy Markov chain, the FuzzRank vector, $\boldsymbol{\mu} = (\mu(1), \ldots, \mu(N))$, may be computed as the limiting vector obtained from Eq. 2 for large values of n. So, in order to obtain the FuzzRanks, all that is required is to obtain accurate values of the elements of the fuzzy transition matrix.

Whenever a page has a single link to another page, it is assumed that there is no fuzziness present there. This is usually the case when a page, say A, has been split into pagelets which were originally its named sections and a link from a page, say B, had specifically pointed to a named section of A (say $A\#C$). Then, after splitting, the page B points to just a single page representing $A\#C$. Had the link just pointed to A without referring to the intended section, the splitting would involve some fuzziness as to which section is being referred to. In that case, the membership values of the target of the link from B are non-zero for multiple pages representing the original sections of A. The membership values may be determined by considering similarity of the context around the anchor of the link and the potential target regions. Thus, the fuzzy transition matrix may be obtained.

The above definition leads to the notion of FuzzRank, which is based on the idea of PageRank. Similar definitions may be provided incorporating the ideas of hubs and authorities, or the whole class of models proposed in [7].

3.3. Advantages and Limitations

We now discuss some features of the proposed class of fuzzy web surfer models. A list of advantages are listed first, following which we delve upon the shortcomings of such a model.

We observe that theoretically, and intuitively, fuzzy web surfer models have the following merits:

1. Capture fuzziness in page contents: page boundaries may not be apparent all the time, especially when a single large page consists of several pagelets. Moreover, noise in web pages also affects the precise identification of the content of interest to the user.
2. Capture fuzziness in links: a page may contain several outlinks but not all of them may be intended for the same purpose. The reason for their presence may be ease of navigation, leading to advertisements, references, or pointing to authoritative resources. Similarly, a link to a particular page may in reality be actually for just one or two sections or pagelets of a page. These kinds of uncertainty may be better modeled by the proposed methodology.
3. Can take into account fuzzy contexts: context sensitive algorithms depend a lot on the modeling assumptions. For example, the context of a query may not be precisely clear, but the system may have a broad idea about it.
4. Robust computations: this is perhaps, the most emphatic reason for choosing fuzzy web surfer models. The computations in $\max$-$\min$ algebra are more robust to perturbations as compared to usual addition and multiplication operations. There is an example in [1]

that demonstrates the robustness of fuzzy Markov systems in comparison to regular Markov chains. When the entries of the transition matrix are perturbed by small quantities, the effects on the stationary distribution of the regular Markov chains are drastic, whereas, for fuzzy Markov chains, the changes are comparable to the perturbations.

5. Finite convergence: the stationary distribution of fuzzy Markov chains can be computed in finite number of steps, whereas, for regular Markov chains, only an approximation may be found as the convergence may not be achieved in finitely many steps. Existing web surfer models assume that, even though convergence is not attained, the order of the probabilities in the obtained distribution suffices.

We now study the possible limitations of the proposed methodology. It is well known that a Markov chain is ergodic if it is regular. However, in the case of fuzzy Markov chains, no such results are known. So, it is not clear when FuzzRank would actually exist, and even when it does, if it would be independent of the initial state of the process. There is an example in [1] where the rows of the limiting fuzzy transition matrix are distinct, and hence demonstrates the existence of non-ergodic fuzzy Markov chains.

This, however, need not be a limitation as all that it implies is that the final fuzzy distribution of the surfer being on a particular page may not be independent of his initial state. In practice, this may indeed be the case as a surfer starting from one set of pages may, in the long run, behave differently from another one who starts from a different set of web pages. Thus, that the fuzzy Markov chain of web pages being visited is not ergodic may be a blessing in disguise, which may be useful in computing topic sensitive page ranks or for detecting web communities.

4. Future Research and Conclusions

The model proposed in this paper appears to be theoretically very interesting. Various advantages have also been discussed. These need to be confirmed experimentally. We also conjecture that the distinct sets of rows of the limiting transition matrix would correspond to distinct communities of web pages. This again needs to be checked experimentally. Also, some studies need to be conducted for obtaining conditions for guaranteeing the aperiodicity and ergodicity of the fuzzy Markov chain, at least in some special cases.

This is a work in progress, and despite the lack of experimental results, we present it in its preliminary form because of the huge implications it has for the field of Web Intelligence. The contribution in this article is the novel theoretical formulation of fuzzy web surfer models by integrating existing works on web surfer models and fuzzy Markov chains, as well as, discussions on where they are applicable, and what their advantages are. Also, the scope for future research on this topic is vast.

5. Acknowledgment

The first author's doctoral fellowship is funded by INSEAD, France.

References

[1] K. Avrachenkov and E. Sanchez. Fuzzy markov chains and decision-making. *Fuzzy Optimization and Decision Making*, 1(2):143–159, June 2002.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual search engine. Technical report, Stanford University, 1998.

[3] J. J. Buckley. Note on: Convergence of powers of a fuzzy matrix. *Fuzzy Sets and Systems*, 121:363–364, 2001.

[4] J. J. Buckley. *Fuzzy probabilities and fuzzy sets for web planning*. Springer, 2004.

[5] J. J. Buckley and E.Eslami. Fuzzy markov chains: Uncertain probabilities. *Mathware and Soft Computing*, 9(1):33–41, 2002.

[6] S. Chakrabarti, M. M. Joshi, and V. B. Tawde. Enhanced topic distillation using text, markup tags and hyperlinks. In *Research and Development in Information Retrieval*, 2001.

[7] M. Diligenti, M. Gori, and M. Maggini. A unified probabilistic framework for web page scoring systems. *IEEE Transactions on Knowledge and Data Engineering*, 16(1):4–16, January 2004.

[8] T. H. Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge and Data Engineering*, 2003.

[9] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[10] B. L. Narayan and S. K. Pal. Detecting sequences and cycles of web pages. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, Compiégne, France, September 2005. (to appear).

[11] S. K. Pal, B. L. Narayan, and S. Dutta. A web surfer model incorporating topic continuity. *IEEE Transactions on Knowledge and Data Engineering*, 17(5):726–729, May 2005.

[12] L. Ramaswamy, A. Iyengar, L. Liu, and F. Douglis. Automatic fragment detection in dynamic web pages and its impact on caching. *IEEE Transactions on Knowledge and Data Engineering*, 17(5):859–874, May 2005.

[13] M. Richardson and P. Domingos. The intelligent surfer: Probabilistic combination of link and content information in pagerank,. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2002.