Time damping of textual relevance

Anand Kishore
anand@semanticvoid.com

Abstract

In the world of Information retrieval items are ordered as per their relevance to the given query. This relevance is derived from a number of characteristics of the object at hand. One such characteristic, for dynamic content is Time. This paper describes a mathematical construct which computes the relevance of an object with respect to Time.

Introduction

Significant focuses of our search on the web today is dynamic or volatile content. Volatile content refers to content which exist for a given period of time, for example discount offers. Dynamic content refers to content whose interest factor decreases with time, for example news articles. For any content that falls into the above two categories, time is an important characteristic.

Most search engines of today provide two metrics for ordering content – one by relevance (textual) and the second by date. But dynamic or volatile data cannot be ordered by textual relevance or date alone. For example, if a search on Google Web History is ordered only by textual relevance then significantly older articles may turn up as top results. A site browsed/searched by a user in the past few months is more relevant to him than what he browsed/searched for a couple of years back. Similarly, a search on Google Web History when ordered by date is equivalent to browsing the entries by date.

Exponential Decay

For the lifetime of a dynamic/volatile content, its relevance decreases with respect to time. This decay cannot be linear in nature as, depending on the characteristic of the content, it is interesting to a user only for a short time span of its life (like a week or a month) after which its interestingness starts to drop sharply. This decay can be modeled into an exponential function.

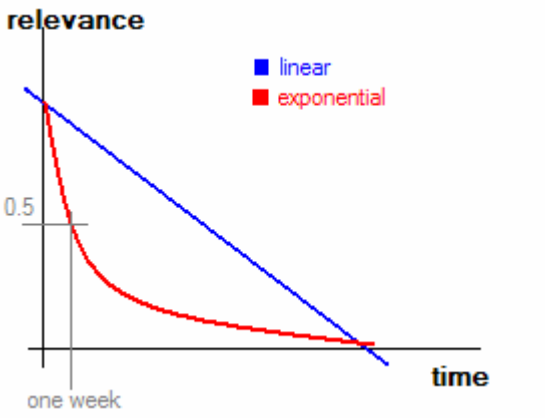

Figure 1

This exponential decrease needs to be quantified such that the overall relevance of an object can be determined as a function of the two factors:

- Time
- Other factors for relevance

Interest Slots

Interest Slots (I) can be defined as the smallest unit of time for which the time relevance (exponential decay) of two

objects is equal. As the object moves from one such slot to another, its time relevance decreases.

For example, let us consider a news article which was indexed today. Assuming the interest slot for news articles to be of 1 day, if the article remains in the index for 30 days its lifetime can be divided into as many such slots which amount to 30 interest slots.

Time Damper

The exponential in figure 1 represents the output of the time damper. As stated above, depending on the characteristics of the object, we can state that the relevance of an object would reduce to x% in time t_x.

Thus the time damping function (d) for such an object at time t would be defined as,

$$d(t,t_0) = k^{dT/tk}$$

where,

t is the current time

t_0 is the time when the object was created (or rather indexed)

k is the known percentage decrease in the relevance (x%)

I_s is the interest slot for the object

dT is time in interest slots elapsed since the object was created, $(t - t_0)/I_s$

t_k is the time in interest slots such that the time damping k is achieved.

By tweaking this factor, we can alter the exponential plot to suit the nature of objects in our system.

This factor when multiplied to the relevance decreases it exponentially.

Example

Let us consider the example where:
I_s = 1 day t_0 = 0
k = 50% t_k = 6 days
= 0.5 = 6 interest slots
relevance (r) = 50 units

The following table lists the effect of time damper on relevance:

t	$d(t,t_0)$	$r = r \times d(t,t_0)$
1	0.890	44.5
2	0.793	39.65
3	0.707	35.35
4	0.629	31.45
5	0.561	28.05
6	0.5	25
7	0.445	22.25
8	0.396	19.8
9	0.353	17.65
10	0.314	15.7
11	0.280	14
12	0.250	12.5
13	0.222	11.1
14	0.198	9.9
15	0.176	8.8
16	0.157	7.85
17	0.140	7
18	0.125	6.25
19	0.111	5.55
20	0.099	4.95
21	0.088	4.4
22	0.078	3.9
23	0.070	3.5
24	0.062	3.1
25	0.055	2.75
26	0.049	2.45
27	0.044	2.2
28	0.039	1.95
29	0.035	1.75
30	0,031	1.55

Figure 2

This graph plots the relevance score versus the time with the values as in the above table.