Category Jaccard averaged on the Outlinks

Jacopo Notarstefano
jacopo.notarstefano [at] gmail.com

11 February 2014

Category Jaccard

Definition (Category Jaccard)

Let p_1 and p_2 be two pages and let C_1 and C_2 be, respectively, their list of categories. Then their Category Jaccard similarity is

$$J_C(p_1, p_2) = \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|}.$$

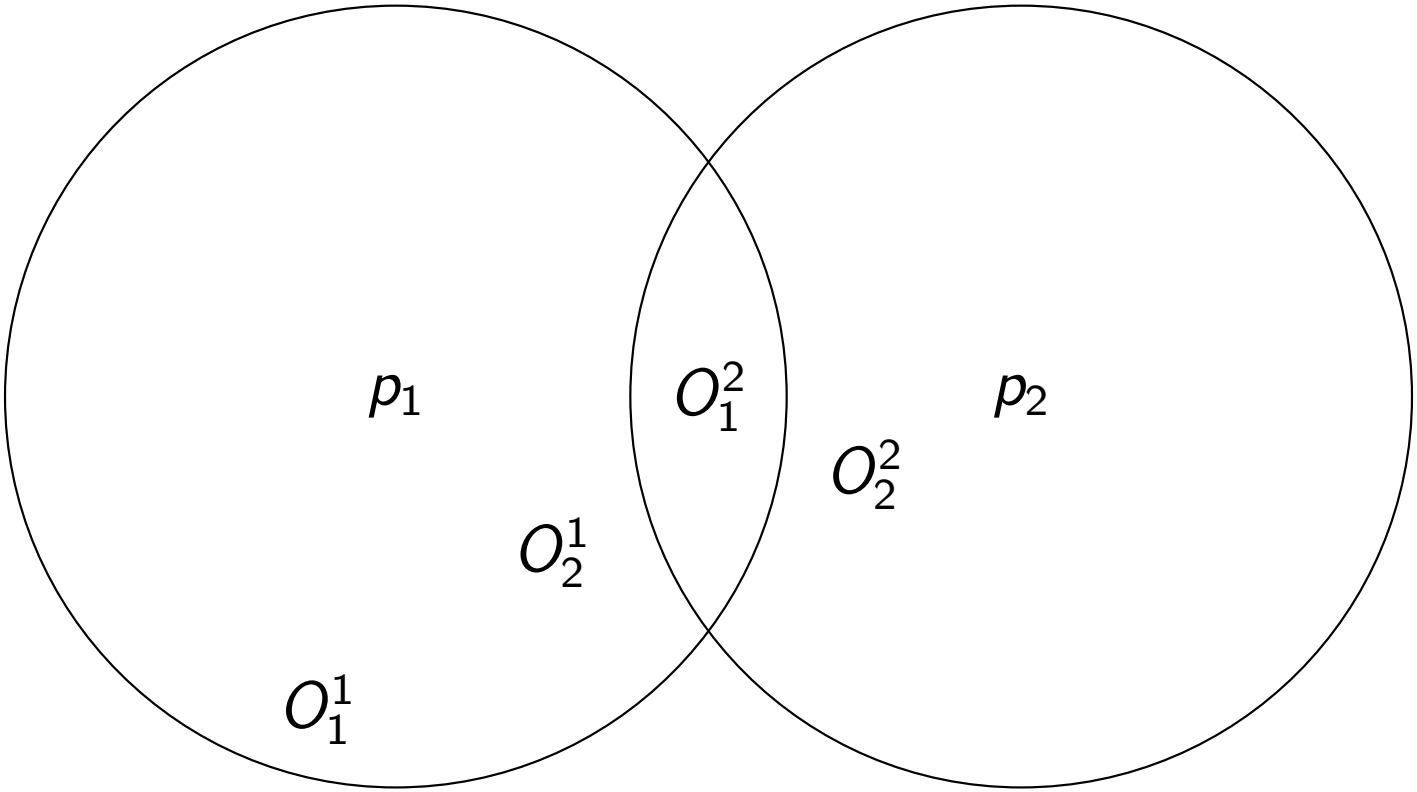
Category Jaccard averaged on the Outlinks

Definition (Category Jaccard averaged on the Outlinks)

Let p_1 and p_2 be two pages and let O^1 and O^2 be, respectively, their list of linked pages. Then their Category Jaccard similarity averaged on the Outlinks is

$$\mathrm{AVG}_{J_C}(p_1, p_2) = \frac{\sum J_C(O_i^1, O_j^2)}{|O^1| \cdot |O^2|}.$$

A picture $\approx 2^{10}$ words



Advantages

Drawbacks